Online Supplement for “Endogenous Returns to Scale”

by Alexandr Kopytov, Mathieu Taschereau-Dumouchel and Zebang Xu

Not for publication

This document is an online supplement for the paper “Endogenous Returns to Scale”. Part A provides the details of the data exercises of Section 6 along with several robustness checks. Part B provides additional information about the calibration exercise of Section 7. Part C provides the proofs of the formal results. Part D contains several extensions of the model along with robustness exercises.

A Supplement for Section 6

This online supplement contains details about the reduced-form results of Section 6.

A.1 Details of the Spanish Orbis data

Our Spanish firm-level data are drawn from the Orbis Historical Disk. Orbis is commonly regarded as the most comprehensive cross-country firm database, covering both public and private firms' financial statements and measures of real activity (Kalemli-Özcan et al., 2024). We focus on Spain because firm coverage is close to universal—capturing over 95% of total industry gross output after 2010—making it well-suited for economy-wide analysis. Our sample spans 1995–2019.[38]

Sample cleaning Our sample construction closely mirrors the cleaning steps used in our earlier work (Kopytov et al., 2024). We begin by merging each firm's descriptive information with its financial statements using the unique BVD firm identifier (`BVDID`). We then restrict our analysis to Spanish firms, defined as firms that satisfy two criteria: 1) their latest address is in Spain and 2) their `BVDID` starts with the ISO-2 code `ES`. In the resulting Orbis Spain sample, we implement the following standard cleaning procedure:

[38]Orbis offers good coverage of the Spanish economy starting from 1995. Moreover, the most recent observations in the version of Orbis Historical Disk Product that we use are from 2021. We therefore use 2019 as the last year of the sample since there is usually a two-year reporting lag for some variables (see Kalemli-Özcan et al. (2024) for details).

1. We harmonize the calendar year of each firm-year observation using the variable `closing_date`: if the closing date is on or after July 1, the current year is assigned as the calendar year. Otherwise, the previous year is assigned.[39]

2. In a given year, a firm in the Orbis database might have multiple financial statements from different sources (local registry, annual report, or others), for consolidated or unconsolidated accounts. When several source-consolidation combinations exist for a firm, we deduplicate by selecting the account that, in order of priority, 1) shows the most consistent reporting frequency (closest to regular annual reporting), 2) offers the longest non-missing time series for key financial variables (fixed assets and/or sales), and 3) is consolidated, if the first two criteria are tied.

3. We only keep firm-year observations with non-missing and positive sales (`operating_revenue_turnover`), fixed assets (`fixed_assets`), wage bills (`costs_of_employees`), and material costs (`material_costs`). We also harmonize the units of all monetary values to be in current euros.

4. To prevent outliers from affecting the production function estimation, we exclude any firm-year observation whose average revenue product of any input (fixed assets, wage bills, or material costs) lies above the 99th percentile or below the 1st percentile of that year's distribution.[40]

A.2 Details of the production function and RTS estimation

This online supplement describes in detail how we implement the production-function estimation procedure that delivers the results used in the main text.[41]

[39]This adjustment matters little for the Spanish sample, as 99% of firms close their books on December 31.

[40]One might worry that trimming on average revenue products of inputs could mechanically remove observations corresponding to extreme returns to scale, since average revenue products of variable inputs are inversely related to the chosen η. In practice, however, the extreme tails in the data are unlikely to reflect meaningful limits of η; they typically coincide with implausible or mismeasured inputs and would require implausibly extreme η to rationalize.

[41]Many recent papers have used production-function estimation to recover heterogeneity in returns to scale at the firm or industry level, including De Loecker et al. (2020), Ruzic and Ho (2023), Chiavari (2024), McAdam et al. (2024), Savagar and Kariel (2024), Demirer (2025), Hubmer et al. (2025) and Gao and Kehrig (2025).

We use the Blundell and Bond (2000) IV-GMM estimator to estimate the production functions as our benchmark. This estimator is designed for dynamic panel settings with persistent firm-level variables and, under standard moment conditions, delivers consistent estimates of output elasticities. Our model imposes a competitive output market in a sector. In this setting, the identifying assumptions are most plausible when there is sufficient persistent variation in predetermined inputs and in the cost of flexible inputs, so that observed input choices are not collinear with unobserved productivity. Recent work by De Ridder et al. (2022) further shows, through Monte Carlo simulations, that this approach performs well when such identifying variation is strong.

Our empirical strategy builds on the model's implication that, within a sector, firms that are similar in size should operate under similar production technologies and therefore exhibit similar returns to scale. We use this prediction to estimate returns to scale across the firm-size distribution. For each sector i and year t, we rank firms by a smoothed measure of size: the 7-year moving average of firm-level log sales computed over the window from $t-3$ to $t+3$. [42] We then assign firms to 10 deciles, $d_t = \{1, \ldots, 10\}$, based on this sector-year ranking.[43] Using a moving average reduces the influence of short-run fluctuations and measurement error in annual sales, and thus provides a more stable proxy for the scale of a firm's production. For each sector-decile-year cell (i, d_t, t) (using the 7-year rolling sample around t), we assume firms share a common Cobb-Douglas technology:

$$
\begin{aligned}
q_{ilt} &= \beta^L_{i,d_t(l),t}\, l_{ilt} + \beta^K_{i,d_t(l),t}\, k_{ilt} + \beta^M_{i,d_t(l),t}\, m_{ilt} + \gamma_t^{i,d_t(l)} + \kappa_{il}^{i,d_t(l)} + a_{ilt}^{i,d_t(l)}, \\
a_{ilt}^{i,d_t(l)} &= \rho^{i,d_t(l)}\, a_{il,t-1}^{i,d_t(l)} + e_{ilt}^{i,d_t(l)}, \qquad |\rho^{i,d_t(l)}| < 1, \\
e_{ilt}^{i,d_t(l)} &\sim MA(0),
\end{aligned}
$$

where q_{ilt}, l_{ilt}, k_{ilt}, m_{ilt} are the logs of output, labor, capital, material inputs for firm l. These

[42] We use the centered window whenever feasible. Near the sample boundaries and when firm-year observations are missing, we use the longest available window to preserve coverage.

[43] We reassign a few small sectors with few firms to closely related sectors that produce similar goods or services, for the purpose of production function estimation. Specifically, (i) we merge sectors 5, 6, 7, and 9—Manufacture of food products, beverages and tobacco products; Manufacture of textiles, wearing apparel and leather products; Manufacture of wood and of products of wood and cork (except furniture); manufacture of articles of straw and plaiting materials; and Printing and reproduction of recorded media—into sector 8 (Manufacture of paper and paper products). (ii) We merge sector 12—Manufacture of basic pharmaceutical products and pharmaceutical preparations—into sector 11 (Manufacture of chemicals and chemical products). (iii) We merge sector 20—Manufacture of motor vehicles, trailers and semi-trailers—into sector 19 (Manufacture of machinery and equipment n.e.c.).

are measured as deflated values of, respectively, sales, wage bills, fixed assets and material costs using the GDP deflator in the Annual Spanish National Accounts. We assume a firm's productivity contains three components: a year-specific component $\gamma_t^{i,d_t(l)}$, a firm-specific effect $\kappa_{il}^{i,d_t(l)}$ and an autoregressive component $a_{ilt}^{i,d_t(l)}$ with i.i.d. innovation $e_{ilt}^{i,d_t(l)}$. The model admits the following dynamic representation:

$$\begin{aligned} q_{ilt} = \rho^{i,d_t(l)} q_{il,t-1} + \beta^L_{i,d_t(l),t}\, l_{ilt} - \rho^{i,d_t(l)} \beta^L_{i,d_t(l),t}\, l_{il,t-1} + \beta^K_{i,d_t(l),t}\, k_{ilt} - \rho^{i,d_t(l)} \beta^K_{i,d_t(l),t}\, k_{il,t-1} \\ + \beta^M_{i,d_t(l),t}\, m_{ilt} - \rho^{i,d_t(l)} \beta^M_{i,d_t(l),t}\, m_{il,t-1} + \gamma_t^{*\,i,d_t(l)} + \kappa_{il}^{*\,i,d_t(l)} + e_{ilt}^{i,d_t(l)}, \end{aligned} \tag{38}$$

where $\gamma_t^{*\,i,d_t(l)} := \gamma_t^{i,d_t(l)} - \rho^{i,d_t(l)} \gamma_{t-1}^{i,d_t(l)}$ and $\kappa_{il}^{*\,i,d_t(l)} := (1-\rho^{i,d_t(l)})\kappa_{il}^{i,d_t(l)}$. Therefore, we can estimate the following dynamic specification with current and lagged variables:

$$\begin{aligned} q_{ilt} = \alpha^{i,d_t(l)}\, q_{il,t-1} + \beta^{L0}_{i,d_t(l),t}\, l_{ilt} + \beta^{L1}_{i,d_t(l),t}\, l_{il,t-1} + \beta^{K0}_{i,d_t(l),t}\, k_{ilt} + \beta^{K1}_{i,d_t(l),t}\, k_{il,t-1} \\ + \beta^{M0}_{i,d_t(l),t}\, m_{ilt} + \beta^{M1}_{i,d_t(l),t}\, m_{il,t-1} + \gamma_t^{*\,i,d_t(l)} + \kappa_{il}^{*\,i,d_t(l)} + e_{ilt}^{i,d_t(l)}, \end{aligned} \tag{39}$$

where, under our assumption, the AR(1) productivity structure implies restrictions across coefficients (e.g., $\alpha^{i,d_t(l)} = \rho^{i,d_t(l)}$ and $\beta^{x1}_{i,d_t(l),t} = -\rho^{i,d_t(l)} \beta^{x0}_{i,d_t(l),t}$ for $x \in \{L, K, M\}$).

Blundell-Bond system-GMM moments We now describe the system-GMM moment conditions we exploited to estimate the model in (39). Our choice of moment conditions follows the exact implementation in Table III, column 5 of Blundell and Bond (2000), where we treat $\{q, l, k, m\}$ as potentially endogenous and use two sets of moments:

(i) Difference equation (levels dated $t-2$ and earlier):

$$\mathbb{E}\Big[x_{il,t-s}\ \Delta e_{ilt}^{i,d_t(l)}\Big] = 0 \quad \text{for } x \in \{q, l, k, m\} \text{ and } s \geq 2. \tag{40}$$

(ii) Levels equation (first differences dated $t-1$ only):

$$\mathbb{E}\Big[\Delta x_{il,t-1}\ \big(\kappa_{il}^{*i,d_t(l)} + e_{ilt}^{i,d_t(l)}\big)\Big] = 0 \quad \text{for } x \in \{q, l, k, m\}\,. \tag{41}$$

Moreover, year dummies are included as controls and treated as exogenous instruments in the levels equation. We implement this estimation using the `xtabond2` command in Stata.

Obtaining the Firm-level Returns-to-Scale Estimates After estimating (39), we then use the minimum distance estimator by Söderbom (2009) to impose the AR(1)-implied restrictions and get the restricted parameter estimates $\left(\hat{\rho}^{i,d_t(l)}, \hat{\beta}^K_{i,d_t(l),t}, \hat{\beta}^L_{i,d_t(l),t}, \hat{\beta}^M_{i,d_t(l),t}\right)$. The estimated returns to scale η_{ilt} for a firm l in sector i and year t is therefore given by the sum of these elasticities:

$$\eta_{ilt} = \hat{\beta}^K_{i,d_t(l),t} + \hat{\beta}^L_{i,d_t(l),t} + \hat{\beta}^M_{i,d_t(l),t}.$$

Because 1995 and 1996 contain relatively few firm-year observations for production-function estimation, we report results using only the 1997-2019 estimates matched to the firm-level data for our analysis.

A.3 Constructing the Törnqvist productivity index

To compare productivity across firms, we rely on the Törnqvist productivity index. We provide here theoretical results about that index to link our model with our estimation procedure.

Lemma 3 shows that returns to scale are increasing in ε_{il}, but we do not observe ε_{il} directly in the data. In addition, comparing measured productivity $e^{\varepsilon_{il}} A_i(\eta_{il}) \zeta(\eta_{il})$ across firms with different technologies faces well-known issues about the choice of units. When going to the data, we rely instead on a Törnqvist productivity index, which is commonly used to compare productivities across firms or countries with different production functions (Caves et al., 1982a; Caves et al., 1982b) and recently in Penn World Table by Feenstra et al. (2015). Specifically, we use the multilateral Törnqvist productivity index by Caves et al. (1982a) that has been extensively used in the firm dynamics context (Aw et al., 2001).

Definition 3 (Multilateral Törnqvist productivity index)**.** Consider a sector i in year t. Let N_{it} be the number of firms observed in (i,t). Define the sector-year reference firm's moments as $\overline{\log Q_{it}} = \frac{1}{N_{it}} \sum_l \log Q_{ilt}$, $\overline{\log O_{it}} = \frac{1}{N_{it}} \sum_l \log O_{ilt}$, $\overline{\beta_{O,it}} = \frac{1}{N_{it}} \sum_l \beta_{O,ilt}$ where $O \in \{K, L, M\}$ and $\beta_{O,ilt}$ are firm-level output elasticity of input O. The multilateral Törnqvist productivity index of firm l is defined as:

$$z_{ilt} := \left(\log Q_{ilt} - \overline{\log Q_{it}}\right) - \sum_{O \in \{K,L,M\}} \frac{1}{2}\left(\beta_{O,ilt} + \overline{\beta_{O,it}}\right)\left(\log O_{ilt} - \overline{\log O_{it}}\right).$$

For any two firms k and l in sector i and year t, we say firm k is more productive than firm l if $z_{ikt} > z_{ilt}$.

Intuitively, the measure z_{ilt} compares productivity between firm l and the reference firm by looking at how much more output one produces relative to the other, adjusting for differences in technology and input use. It is "multilateral" because z_{ilt} is defined relative to a common sector-year reference firm constructed from *all firms* in (i,t), so productivity comparisons $z_{ikt} - z_{ilt}$ are base-firm invariant and can be consistently ranked across all firm pairs. In our benchmark case, we set all $\beta_{O,ilt} = \hat{\beta}^{O}_{i,d_t(l),t}$ to obtain the estimated productivity index $\hat{z}_{ilt}$ and use it as our measured productivity in all cross-sectional exercises that involve comparisons between firms within a sector-year. We find the Törnqvist index to be a good proxy for productivity in model-simulated data. In our calibrated economy of Section 7, the within-sector correlation between the Törnqvist index and ε_{il} is above 0.99. The same number for $\varepsilon_{il} + a_i(\eta_{il})$ is about 0.92.

However, when analyzing within-firm productivity changes over time, we use a chained (within-firm) Törnqvist productivity index—i.e., an approximate Divisia index—following the implementation in Star and Hall (1976). Specifically, to account for the fact that firms may simultaneously adjust both their technology (and hence elasticities) and their input mix, we define

$$\Delta\hat{z}^{\text{within}}_{ilt} = \Delta\log Q_{ilt} - \sum_{O\in\{K,L,M\}} \overline{\beta_{O,ilt}}\Delta\log O_{ilt}, \quad \text{where } \overline{\beta_{O,ilt}} \equiv \frac{1}{2}\left(\hat{\beta}^{O}_{i,d_t(l),t} + \hat{\beta}^{O}_{i,d_{t-1}(l),t-1}\right).$$

We then normalize each firm's initial (log) within-firm productivity to zero and construct the level index $\hat{z}^{\text{within}}_{ilt}$ by accumulating changes over time, i.e.,

$$\hat{z}^{\text{within}}_{il,t} = \hat{z}^{\text{within}}_{il,t-1} + \Delta\hat{z}^{\text{within}}_{ilt}, \quad \hat{z}^{\text{within}}_{il,t_0} = 0.$$

This normalization is innocuous because our within-firm analysis in 6.3.2 includes firm fixed effects, so only productivity changes (not the level) are identified.

A.4 Robustness of the production function and RTS estimation

This online supplement shows that the documented positive RTS-size and RTS-productivity relationships, both across firms in the cross-section (Section 6.3.1) and within a firm over time (Section 6.3.2), are not driven by a particular estimator, choice of IV-GMM instruments, or grouping design. We first vary the Blundell and Bond (2000) system-GMM specification by changing the treatment of year dummies and the internal instrument set, following the implementation in De

Ridder et al. (2022) (Supplement A.4.1). We then re-estimate production functions using standard control-function approaches–Olley and Pakes (1996) and Levinsohn and Petrin (2003)–to verify that our results are not specific to IV-GMM (Supplement A.4.2). Moreover, we account for potential market power by adding markup controls (proxied by sales shares) within an Ackerberg et al. (2015) estimator (Supplement A.4.3). Finally, we show that our conclusions are robust to alternative ways of forming size groups (Supplement A.4.4).

We report coefficients from simple regressions to summarize the robustness of our empirical findings both across firms and within firms with these alternative estimates of returns to scale and productivity. To show robustness for Figure 7, which documents the cross-sectional pattern that larger and more productive firms have higher returns to scale within a sector-year, we estimate two simple regressions of returns to scale on log sales and productivity:

$$\eta_{ilt} = \beta_0 \log(\text{Sales}_{ilt}) + \delta_{it} + \epsilon_{ilt}, \qquad \eta_{ilt} = \beta_1 \hat{z}_{ilt} + \delta_{it} + \epsilon_{ilt}, \tag{42}$$

where δ_{it} denotes sector-year fixed effects. The estimated coefficients of β_0 and β_1 are displayed in Table 2 across all specifications.

Similarly, to show robustness for Figure 8 and document our within-firm pattern that firms have higher returns to scale when they grow larger or become more productive, we estimate:

$$\eta_{ilt} = \gamma_0 \log(\text{Sales}_{ilt}) + \kappa_{il} + \delta_{it} + \epsilon_{ilt}, \qquad \eta_{ilt} = \gamma_1 \hat{z}^{\text{within}}_{ilt} + \kappa_{il} + \delta_{it} + \epsilon_{ilt}, \tag{43}$$

where κ_{il} denotes firm fixed effects, so identification comes from within-firm variation over time. In the productivity specification, we use a within-firm Törnqvist productivity index, $\hat{z}^{\text{within}}_{ilt}$, which is appropriate for within-firm comparisons.[44] The estimated coefficients of γ_0 and γ_1 are presented in Table 3 across all specifications.

Results using our benchmark estimator are reported in column (1) of Tables 2 and 3. We now describe the alternative estimators and grouping designs used in the robustness checks.

[44] Since the within-firm productivity index $\hat{z}^{\text{within}}_{ilt}$ is a chained index, it can be constructed only for firm-years with complete sales and input data from the firm's first observation onward (so that the chain can be formed), which reduces the usable sample.

Table 2: Across-firm variation in returns to scale, productivity, and firm size with alternative production function estimators

	Dependent variable: Firm-level RTS						
	(1) BB baseline	(2) BB alternative	(3) OP	(4) LP	(5) ACF market power	(6) Av.-size percentiles	(7) Cur.-size deciles
$\log(\text{Sales}_{ilt})$	0.023***	0.028***	0.045***	0.037***	0.050***	0.019***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952
R^2	0.688	0.688	0.642	0.806	0.564	0.728	0.655
$\hat{z}_{ilt}$	0.050***	0.058***	0.083***	0.080***	0.091***	0.034***	0.021***
	(0.002)	(0.003)	(0.002)	(0.003)	(0.005)	(0.002)	(0.002)
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952	9,424,952
R^2	0.655	0.648	0.567	0.760	0.507	0.684	0.652

Notes: This table reports coefficients from cross-sectional regressions of firm-level returns to scale (RTS) on (i) log sales and (ii) firm productivity, $\widehat{z}_{ilt}$, each including sector-year fixed effects. Column (1) uses our benchmark Blundell–Bond (BB) estimates; column (2) uses an alternative BB specification as implemented in De Ridder et al. (2022). Columns (3) and (4) use the Olley–Pakes (OP) and Levinsohn–Petrin (LP) control-function estimators, respectively. Column (5) reports results from the Ackerberg–Caves–Frazer (ACF) estimator with market power controls (proxied by firms' sales shares). Columns (6) and (7) use alternative grouping methods for estimating elasticities: rolling average-size percentiles and contemporaneous size deciles. The regressions use a sample of Spanish firms from Orbis. See Supplement A.1 for details on variable construction and sample selection. Standard errors (in parentheses) are two-way clustered at the firm and sector-year level. *,** ,*** indicate significance at the 10%, 5%, and 1% levels, respectively.

A.4.1 With alternative Blundell-Bond specifications

Our baseline specification follows Blundell and Bond (2000) and includes year dummies. Including year effects is recommended in dynamic-panel GMM applications because it absorbs economy-wide shocks and thereby reduces cross-firm correlation in the regression residuals. At the same time, once common year shocks are removed, identification of flexible-input elasticities relies on variation that is not common across firms in a group. In practice, this shifts weight toward persistent within-year differences in flexible input costs or wedges across firms. If such variation is interpreted literally as firm-specific input prices, it can raise concerns about measurement—because input quantities constructed from expenditures may mechanically inherit noise from unobserved firm-level prices. [45]

As a robustness check, we therefore also implement the Blundell–Bond estimator specification

[45] However, if the relevant heterogeneity operates through non-monetary wedges—e.g., distortions that affect effective input costs without changing the recorded unit prices paid by the firm, in the spirit of Hsieh and Klenow (2009)—then this concern is mitigated because observed input quantities are not mechanically distorted by unobserved prices.

Table 3: Within-firm variation in returns to scale, productivity, and firm size with alternative production function estimators

	Dependent variable: Firm-level RTS						
	(1) BB baseline	(2) BB alternative	(3) OP	(4) LP	(5) ACF market power	(6) Av.-size percentiles	(7) Cur.-size deciles
$\log(\text{Sales}_{ilt})$	0.013***	0.018***	0.022***	0.019***	0.027***	0.010***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,248,461	9,248,461	9,248,461	9,248,461	9,248,461	9,248,461	9,248,461
R^2	0.799	0.813	0.875	0.927	0.693	0.853	0.739
$\hat{z}_{ilt}^{\text{within}}$	0.008***	0.008***	0.012***	0.014***	0.021***	0.005***	0.004***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sector-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,254,839	5,254,839	5,254,839	5,254,839	5,254,839	5,254,839	5,254,839
R^2	0.829	0.845	0.911	0.947	0.741	0.882	0.768

Notes: This table reports coefficients from within-firm regressions of firm-level returns to scale (RTS) on (i) log sales and (ii) within-firm productivity, $\hat{z}_{ilt}^{\text{within}}$, each including firm fixed effects and sector-year fixed effects. Column (1) uses our benchmark Blundell–Bond (BB) estimates; column (2) uses an alternative BB specification as implemented in De Ridder et al. (2022). Columns (3) and (4) use the Olley–Pakes (OP) and Levinsohn–Petrin (LP) control-function estimators, respectively. Column (5) reports results from the Ackerberg–Caves–Frazer (ACF) estimator with market power controls (proxied by firms' sales shares). Columns (6) and (7) use alternative grouping methods for estimating elasticities: rolling average-size percentiles and contemporaneous size deciles. The regressions use a sample of Spanish firms from Orbis. See Supplement A.1 for details on variable construction and sample selection. Standard errors (in parentheses) are two-way clustered at the firm and sector-year level. *,**,*** indicate significance at the 10%, 5%, and 1% levels, respectively.

used by De Ridder et al. (2022), which omits year dummies and uses a more conservative internal-instrument set. Concretely, our baseline estimates a dynamic sales equation with current and one-lag terms for labor, capital, and materials, includes year fixed effects, and instruments the endogenous variables with lags starting at $t-2$ (and deeper) in the transformed equation, while treating the year dummies as standard instruments in the levels equation. In contrast, the De Ridder et al. (2022) specification removes year dummies and restricts the GMM-style instruments to a single deeper lag (the third lag) for output and inputs. Relative to our baseline, this alternative places less weight on within-year cross-sectional price/wedge variation as the driver of instrument relevance and also reduces instrument proliferation by construction. The results using the alternative Blundell-Bond estimates are reported in column (2) of Tables 2 and 3.

A.4.2 With different production function estimators

We also use other commonly used production function estimators as a robustness check. In particular, we consider the control-function approach and implement the Olley and Pakes (1996) and Levinsohn and Petrin (2003) estimators.

The Olley–Pakes Estimator We first implement the Olley and Pakes (1996) (OP) estimator to estimate the production function for each sector–decile–year cell. The Olley and Pakes (1996) estimator is a semiparametric control-function method that addresses simultaneity bias, since unobserved productivity affects firms' input choices. It assumes that investment is a function of capital and productivity and, under a monotonicity condition, can be inverted to express unobserved productivity in terms of observed investment and capital. Substituting this inverted control function into the production function, the method first estimates the elasticities of freely adjustable inputs (labor and materials in our case) while controlling for productivity, and then uses a Markov assumption on productivity to recover the coefficient on the quasi-fixed input, capital. To implement this approach, we measure real investment as the change in the capital stock net of depreciation, and we recognize that this can generate zero or negative investment values, which reduces the usable sample for production function estimation.[46] The results using the OP estimator are reported in column (3) of Tables 2 and 3.

The Levinsohn–Petrin Estimator Because investment can be lumpy in practice and the Olley and Pakes (1996) procedure may force us to drop observations with zero or negative investment, we also apply the Levinsohn and Petrin (2003) (LP) estimator as an additional robustness check. Instead of using investment, this method uses intermediate inputs (materials in our case) as the control variable. It assumes that materials are flexibly chosen after observing productivity, while capital is still treated as quasi-fixed. Under the assumption that materials demand is a function of capital and productivity and is monotone in productivity (conditional on capital), the materials demand function can be inverted to recover unobserved productivity. This control function allows consistent estimation of the labor elasticity, and additional moment conditions then recover the elasticities of capital and materials. The results using the LP estimator are reported in column (4) of Tables 2 and 3.

[46] This reduces the production function estimation sample within each sector–decile–year cell, but we apply the resulting coefficients to the common baseline panel, so the main regression sample is unchanged.

A.4.3 With controls for market power

Our model abstracts from market power and markups, but these forces could hinder the identification of output elasticities and obscure the positive RTS–size relationship we identify in the cross section. When firms have market power, they may charge different output prices, so elasticities estimated using deflated sales can be closer to revenue elasticities than to physical output elasticities.

That said, we do not expect this channel to explain our results. Under monopolistic competition, larger firms typically have higher markups. Higher markups mechanically dampen the sensitivity of revenue to input expansion, implying lower revenue elasticities for larger firms relative to smaller firms. If anything, this would bias against finding a positive RTS–size relationship. Therefore, the presence of markups would tend to weaken our estimated positive RTS–size relationship, suggesting that the underlying relationship could be even stronger.

Nonetheless, we follow common practice and re-estimate the production function with explicit controls for market power, treating price variation as an additional measurement component to be partialed out. Specifically, following Baqaee and Farhi (2019) and De Loecker et al. (2020), we control for markups using firms' sales shares (measured at the NACE 3-digit and 4-digit levels) and estimate production functions using the Ackerberg et al. (2015) (ACF) estimator. The results using the ACF estimator are reported in column (5) of Tables 2 and 3.

A.4.4 With different size-based grouping methods

Grouping firms by 7-year average sales percentiles Our benchmark approach groups firms in sector i and year t into deciles based on their 7-year average log sales. While straightforward, this discretization can generate non-smooth variation across firm sizes. As a robustness check, we therefore implement a rolling-percentile approach based on firms' 7-year average sales. For each sector-year, we rank firms into 100 percentiles using their 7-year average (log) sales. In each sector i, for each percentile p_t, we construct a local sample consisting of firms whose percentile rank lies between $p_t - 15$ and $p_t + 15$ in year t. We then estimate output elasticities for each cell (i, t, p_t) using the Blundell–Bond estimator on the corresponding 7-year rolling-window sample. The results using the rolling-percentile grouping approach are reported in column (6) of Tables 2 and 3.

Grouping firms by contemporaneous sales deciles Alternatively, we group firms into deciles based on contemporaneous firm-level (log) sales in year t, rather than the 7-year average. We then estimate output elasticities for each cell (i, t, d_t) using the same Blundell–Bond estimator on a 7-year rolling-window sample. The results using the contemporaneous sales-decile grouping approach are reported in column (7) of Tables 2 and 3.

A.4.5 Summary

Overall, the main empirical patterns are robust. Across all alternative production-function estimators (alternative Blundell-Bond specifications, Olley-Pakes, Levinsohn-Petrin, and ACF with market-power controls) and alternative grouping methods (rolling percentiles and contemporaneous deciles), we continue to find a positive relationship between firm-level returns to scale and firm size, as well as between returns to scale and productivity, both in the cross section (within sector-years) and within firms over time. While magnitudes vary across specifications, the sign and statistical significance of these relationships are stable (see Tables 2 and 3).

A.5 Estimation of the tail index

This online supplement describes how we estimate the tail index of the firm-size distribution in each sector-year using the log-rank estimator of Gabaix and Ibragimov (2011). For each sector i and year t, let S_{ilt} denote firm l's sales, and let N_{it} be the number of firms observed in (i, t). We assign ranks $r = 1, \ldots, N_{it}$ according to their sales, where $r = 1$ corresponds to the firm with the largest sales. Let $S_{i(1)t} \geq S_{i(2)t} \geq \cdots \geq S_{i(N_{it})t}$ denote sales sorted in descending order within sector-year (i, t).

We focus on the right tail of the sales distribution and select the tail sample as follows: If $N_{it} > 5000$, we use firms in the top 1% of the sales distribution in (i, t). If $N_{it} \leq 5000$, we use the 50 firms with the largest sales in (i, t).[47] For each sector-year (i, t), we estimate the Pareto tail index ζ_{it} within the tail sample using the Gabaix and Ibragimov (2011) bias-corrected log-rank regression:

$$\log\left(r - \frac{1}{2}\right) = a_{it} - \zeta_{it} \log S_{i(r)t} + u_{irt}. \tag{44}$$

This regression relates the log bias-corrected rank $\log\left(r - \frac{1}{2}\right)$ to log sales. We recover $\widehat{\zeta}_{it}$ as the negative of the OLS slope coefficient on $\log S_{i(r)t}$ and use it as the tail index of sales in Figure 9.

A.6 Details of the imported-input tariff shock exercise

This online supplement provides additional details on the imported-input tariff shock used in Section 6.3.2. Our goal is to measure changes in input costs driven by changes in import tariffs.

[47] If fewer than 50 firms are observed, we use all available firms.

To isolate variation that differs across downstream sectors and over time, we construct a shift-share exposure measure that combines (i) predetermined import input shares from the OECD multi-country input–output tables and (ii) tariff changes from the Global Tariff Project (Teti, 2024).

Let downstream sectors in Spain be indexed by i. Index a foreign exporter-sector pair by $n = (c, s)$, where c denotes the exporting country and s the exporting sector. For each Spanish downstream sector i and year t, we define the tariff-based input cost shifter as

$$\log T_{it} = \sum_{c,s} \Big(\text{ImportShare}^{\text{Intermediate}}_{(\text{Spain},i)\leftarrow(c,s),t-1} \cdot \log\big(1 + \text{TariffRate}^{\text{Spain}}_{(c,s),t}\big)\Big), \tag{45}$$

where $\text{ImportShare}^{\text{Intermediate}}_{(\text{Spain},i)\leftarrow(c,s),t-1}$ is the share of sector i's total intermediate inputs imported from exporter-sector $n = (c, s)$, measured in year $t-1$ using the OECD multi-country input-output tables. $\text{TariffRate}^{\text{Spain}}_{(c,s),t}$ is the ad valorem tariff rate applied by Spain to imports from exporter-sector (c, s) in year t, taken from the Global Tariff Project. Sector i and foreign sectors s are defined according to the OECD input–output classification, which is slightly more aggregated than the NACE 2-digit level. When tariff data are available at a more disaggregated level in Teti (2024), we aggregate to (c, s) using a simple (unweighted) mean across subsectors. Note that $\log T_{it}$ is essentially a weighted average of log tariff factors across upstream foreign inputs, with weights given by the downstream sector's lagged import input structure. It rises when tariffs increase on inputs that the sector i relies on more intensively. The shift-share structure uses lagged import shares to reduce concerns that contemporaneous changes in sourcing respond mechanically to tariff changes.

We then estimate the dynamic impact of these shocks on returns to scale using panel local projections for horizon years $h = -2, \dots, 5$:

$$\eta_{il,t+h} - \eta_{il,t-1} = \beta_h \log T_{it} + \gamma_{lh} + \gamma_{th} + \varepsilon_{ilth},$$

controlling for firm (γ_{lh}) and year (γ_{th}) fixed effects. Under the assumption that tariff changes for a given exporter-sector pair (c, s) are not systematically correlated with unobserved, time-varying shocks to Spanish downstream sector i (conditional on these fixed effects), variation in $\log T_{it}$ provides plausibly exogenous movements in input costs across sectors and over time.

A.7 Details of the cross-country firm-level data

This online supplement describes the firm-level data sources and sample construction for our cross-country analysis. We augment the analysis with firm-level data from a total of 24 countries (including Spain). For 22 European countries, we use Orbis and restrict attention to countries with good coverage of the variables required for production-function estimation. For developing countries, we use China's National Bureau of Statistics (NBS) manufacturing firm database and India's Annual Survey of Industries (ASI). Both the NBS and ASI datasets are censuses of above-scale manufacturing firms.[48] To ensure comparability across countries, we restrict all datasets to manufacturing firms. For each country, we select a seven-year window that maximizes the number of firm-year observations. We briefly discuss the data cleaning below.

Orbis For Orbis, we start from the raw firm-year panel for each country and apply the same four-step cleaning procedure used in Supplement A.1 for Spain. We then (i) restrict the sample to manufacturing firms (corresponding to USSIC codes 2000-3999) and (ii) deflate all nominal financial variables using the country-specific GDP deflator from the World Bank. After cleaning and deflation, we implement the seven-year window selection described above and keep the window with the largest number of firm-year observations for each country.

India ASI Our Indian data come from the Annual Survey of Industries (ASI) for 1998–2018. We harmonize industry codes to NIC-2004 and then map them to the USSIC division level, retaining only manufacturing divisions. We measure sales using the gross sale value of all products. We measure capital using the average of the opening and closing gross book value of total capital. We measure labor using total wage bills. All variables are deflated using India's GDP deflator from the World Bank. We then select the seven-year window with the largest number of firm-year observations (2012–2018).

China NBS The China data are annual firm-level surveys collected by the National Bureau of Statistics (NBS). We use the 1998–2007 sample period. We measure sales using product sales revenue, capital using total fixed assets, and labor using total annual wages payable. Firms are classified by a four-digit Chinese Industry Classification (CIC) code, which we harmonize to the USSIC division level. We retain manufacturing divisions only. All nominal variables are deflated using China's GDP deflator from the World Bank. We then select the seven-year window with the largest number of firm-year observations within the available sample period (2001–2007).

[48]For China, the NBS data also include state-owned enterprises below these thresholds. For India, the ASI includes a representative sample of small firms below the census cutoff.

Production function and RTS estimation We estimate production functions using the Blundell-Bond approach, following our baseline estimation strategy. We treat manufacturing as a single sector within each country. For each country c, let $[t_m(c)-3, t_m(c)+3]$ denote the selected seven-year window and $t_m(c)$ is the median year of that window. We only estimate production functions for firms existing in the median year $t_m(c)$. We group firms into deciles for year $t_m(c)$ based on their seven-year average log sales. We then estimate a decile-specific Cobb-Douglas production function using the full seven-year panel.

Let $\hat{\beta}^O_{c,d(l),t_m(c)}$ denote the estimated output elasticity of input $O \in \{K, L, M\}$ for country c and sales decile d. The returns to scale assigned to firm l in country c at year $t_m(c)$ is computed as the sum of the estimated input elasticities:

$$\eta_{clt_m(c)} = \hat{\beta}^K_{c,d(l),t_m(c)} + \hat{\beta}^L_{c,d(l),t_m(c)} + \hat{\beta}^M_{c,d(l),t_m(c)}.$$

We then construct the Törnqvist productivity index $\widehat{z}_{clt_m(c)}$ using these estimates and compute the covariance between returns to scale and log sales, as well as between returns to scale and productivity $\widehat{z}_{clt_m(c)}$ used in Figure 10 panel (a). In panel (b), we plot the seven-year average ($t_m(c)-3$ to $t_m(c)+3$) of log GDP per capita obtained from Penn World Table version 11.0 against the covariance between returns to scale and productivity $\widehat{z}_{clt_m(c)}$.

B Supplement for Section 7

This section contains details about the calibration of Section 7.

B.1 Calibration data

This section describes the datasets used in the calibration and how the associated sectoral moments are computed.

1. We calibrate the sectoral parameters using the 2010 input-output table from the Spanish National Accounts. This table partitions the Spanish economy into 62 sectors which are usually defined at the 2-digit NACE industry level.[49] Conforming to the accounting conventions in

[49]Sector 63 (household-related production activities) and sector 64 (services by extraterritorial organizations and bodies) are also present in the 2010 input-output table, but their input-output data is missing.

the data, we calibrate the input elasticities of good s' in the production of sector s as

$$\hat{\alpha}_{ss'} = \frac{\text{Input from } s' \text{ at basic prices}_s}{\text{Total input at basic prices}_s} \times \frac{\text{Intermediate consumption at purchaser's prices}_s}{\text{Intermediate consumption at purchaser's prices}_s + \text{total labor expenditure}_s}$$

and the labor elasticity as

$$1 - \sum_{s'} \hat{\alpha}_{ss'} = \frac{\text{total labor expenditure}_s}{\text{Intermediate consumption at purchaser's prices}_s + \text{total labor expenditure}_s},$$

which corresponds to the labor share of total cost in the data.[50,51] We calibrate the consumption share β_s to be the share of final consumption expenditure of good s in the sum of consumption expenditure spent on the 62 sectors.

2. We compute cross-sectional moments from the Orbis sample. After steps 1-4 in Supplement A.1 and the production function estimation in Supplement A.4.1, we perform a few additional steps:

   (a) We winsorize the estimated returns to scale η_{ilt} at the top or bottom 0.5% of the firm-year distribution. In addition, we cap values above 0.99 at 0.99. Using firm-level returns to scale η_{ilt}, we compute each sector's effective returns to scale $\hat{\eta}_{it}$ as the sales-weighted average of these firm-level estimates.

   (b) We compute profits as $\Pi_{ilt} = (1 - \eta_{ilt}) P_{it} Q_{ilt}$ and winsorize it at the top or bottom 0.5% within each sector–year.

   (c) We then compute the interquartile ranges of Π_{ilt} and η_{ilt} at the sector-year level.

   (d) Finally, we average these sector–year moments over time to obtain sector-level moments used in our static model.

[50] Because sector-to-sector data at purchasers' prices (i.e., adjusted for taxes less subsidies on products) are unavailable, we calibrate the intermediate-input expenditure share of inputs from sector s' used by sector s with the share computed in basic prices, $\frac{\text{Input from } s' \text{ at basic prices}_s}{\text{Total input at basic prices}_s}$.

[51] We deliberately omit capital in this calibration because the Spanish input–output table does not distinguish the user cost of capital from profits.

B.2 Interquartile ranges for returns to scale and profits

From (12), we have

$$\eta_{il} = 1 - \frac{1}{\frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\varepsilon_{il}-\mu_i}{2\gamma_i}},$$

which implies[52]

$$\text{IQR}\left(\eta_{il}\right) = \frac{1}{\frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\sigma_i}{2\gamma_i}\Phi^{-1}\left(0.25\right)} - \frac{1}{\frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\sigma_i}{2\gamma_i}\Phi^{-1}\left(0.75\right)}, \tag{46}$$

where $\Phi\left(\cdot\right)$ is the cumulative distribution function of the standard normal random variable.

Profit of firm l in sector i is given by (54). Plugging (10) and (12) in this expression, we get

$$\log \Pi_{il} = \frac{1}{4\gamma_i}\left(2\gamma_i \frac{1-\varphi_i}{1-\hat{\eta}_i} - \mu_i + \varepsilon_{il}\right)^2 + \log H_i, \tag{47}$$

which implies[53]

$$\text{IQR}\left(\log \Pi_{il}\right) = \frac{\sigma_i^2}{4\gamma_i}\left(F^{-1}_{\chi_1^2\left(\frac{2\gamma_i(1-\varphi_i)}{\sigma_i(1-\hat{\eta}_i)}\right)}\left(0.75\right) - F^{-1}_{\chi_1^2\left(\frac{2\gamma_i(1-\varphi_i)}{\sigma_i(1-\hat{\eta}_i)}\right)}\left(0.25\right)\right), \tag{48}$$

where $F_{\chi_1^2(x)}\left(\cdot\right)$ is the cumulative distribution function of noncentral χ^2 distribution with one degree of freedom and the non-centrality parameter x, and $\varphi_i = \frac{\sigma_i^2}{2\gamma_i}$.

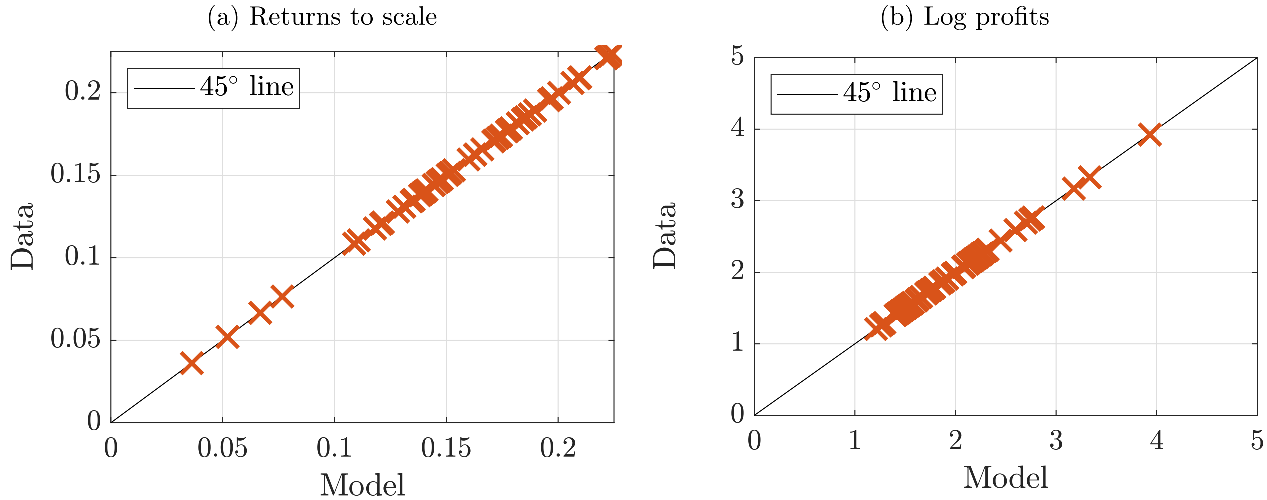
Equations (46) and (48) make clear that IQRs of returns to scale and log profits are functions of σ_i, γ_i, and $\hat{\eta}_i$. We can, therefore, use them to identify σ_i and γ_i. We choose σ_i and γ_i to minimize the distance between model-implied and empirical IQRs, with a constraint $\varphi_i \in [0,1]\ \forall i$. Figure 13 shows that the calibrated model matches the targeted IQRs well.

Figure 14 shows calibrated values of σ_i and γ_i for all sectors. The sector with most volatile productivity is "Petroleum", with $\sigma_i = 3.09$. At the same time, this sector has a high cost of adjusting returns to scale, $\gamma_i = 7.24$, meaning that its effective productivity dispersion is not too large, $\varphi_i = 0.66$.

[52]For all firms with $\eta_{il} \in (0,1)$, η_{il} is strictly increasing in ε_{il}. In the calibrated economy, the fraction of firms with $\eta_{il} \notin (0,1)$ is very small.

[53]From (10) and (12) , $2\gamma_i \frac{1-\varphi_i}{1-\hat{\eta}_i} - \mu_i + \varepsilon_{il} > 0$ for all firms with $\eta_{il} \in (0,1)$. For these firms, $\log \Pi_{il}$ is strictly increasing in ε_{il}. In the calibrated economy, the fraction of firms with $\eta_{il} \notin (0,1)$ is very small.

Figure 13: Interquartile ranges in returns to scale and profits



Notes: Panels (a) and (b) report sectoral interquartile ranges in returns to scale and log profits in the calibrated model and in the data.

B.3 Calibration details for Section 7.3

We analyze the model with sales tax in Supplement D.6. In that supplement, we show that the model with sales taxes can be analyzed analogously to the main model if we properly redefine the mean and the variance of sectoral shocks:

$$\tilde{\mu}_i = \mu_i + \log\left(1 - \tau_i^S\right) \quad \text{and} \quad \tilde{\sigma}_i = \left(1 - b_i\right)\sigma_i.$$

We can identify $\tilde{\sigma}_i$ and γ_i in the same way as described in Supplement B.2. The only difference is that we need to use after-tax profits in (48).

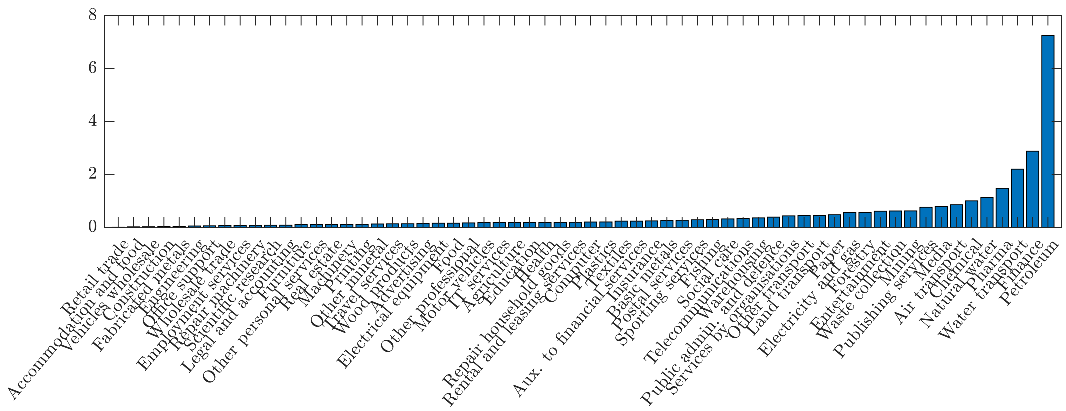
To pin down the parameters of the tax process (37), we proceed as follows. In the data, we compute the covariance of pre-tax profits with $\log\left(1 - \tau_{il}^S\right)$ for each sector. Using (47), we can compute the model analogue of this quantity as

$$\operatorname{Cov}\left(\log \Pi_{il}, \log\left(1 - \tau_{il}^S\right)\right) = \operatorname{Cov}\left(\frac{1}{4\gamma_i}\left(2\gamma_i \frac{1 - \tilde{\varphi}_i}{1 - \hat{\eta}_i} - \tilde{\mu}_i + \tilde{\varepsilon}_{il}\right)^2 - \log\left(1 - \tau_{il}^S\right), \log\left(1 - \tau_{il}^S\right)\right) =$$
$$= -\frac{b_i}{1 - b_i}\tilde{\sigma}_i^2\left(\frac{1 - \tilde{\varphi}_i}{1 - \hat{\eta}_i} + \frac{b_i}{1 - b_i}\right).$$

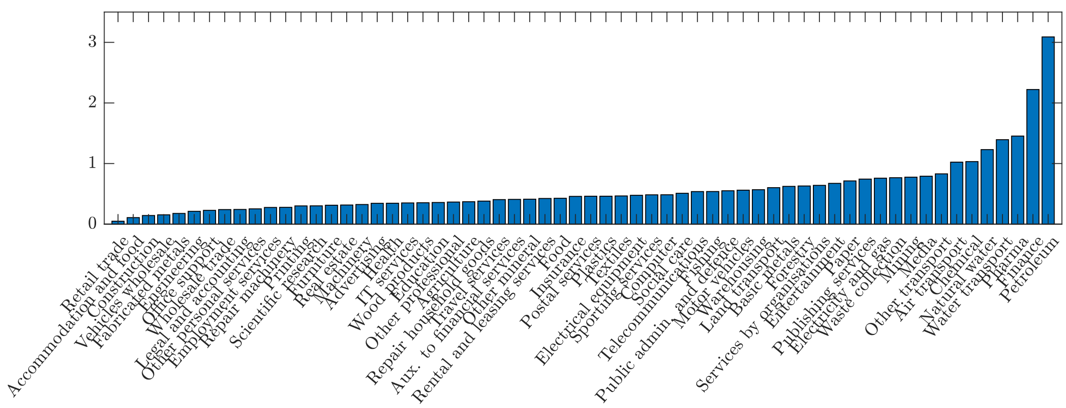
We can identify b_i from this equation.

Figure 14: Calibrated γ_i and σ_i

(a) Calibrated γ_i



(b) Calibrated σ_i



To compute τ_i^S,we rely on equation (87), derived in Supplement D.6, which we reproduce below:

$$\frac{1}{1-\hat{\tau}_i^S} = \frac{1}{1-\tau_i^S}\left(1+\frac{\tilde{\varphi}_i}{\frac{1-\tilde{\varphi}_i}{1-\hat{\eta}_i}}\frac{b_i}{1-b_i}\right)\exp\left(-\frac{b_i}{1-b_i}\frac{4\tilde{\varphi}_i\gamma_i\frac{1-\tilde{\varphi}_i}{1-\hat{\eta}_i}+\tilde{\sigma}_i^2\frac{b_i}{1-b_i}}{2}\frac{1}{1-\tilde{\varphi}_i}\right). \tag{88}$$

In the data, we can observe $\hat{\tau}_i^S$ as the sales-weighted average tax rate in sector i. Then, (87) can be used to identify τ_i^S.

Finally, equation (86) makes clear that the proper measure of sectoral returns to scale $\hat{\eta}_i$ uses after-tax sales as weights.

B.4 Additional quantitative results

Figure (15) shows effective sectoral returns to scale $\hat{\eta}_i$ for all sectors. In our data, the sector with lowest returns to scale is "Water transport" with $\hat{\eta}_i = 0.54$, and the sector with the highest returns to scale is "Retail trade" with $\hat{\eta}_i = 0.98$. The mean and median returns to scale are both 0.83 and 0.82, respectively.

Figure 15: Effective returns to scale $\hat{\eta}_i$ across sectors

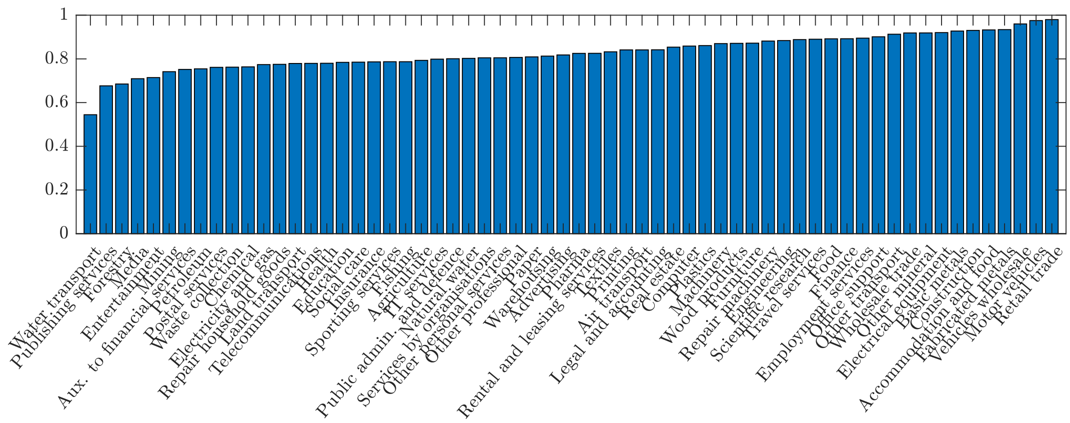
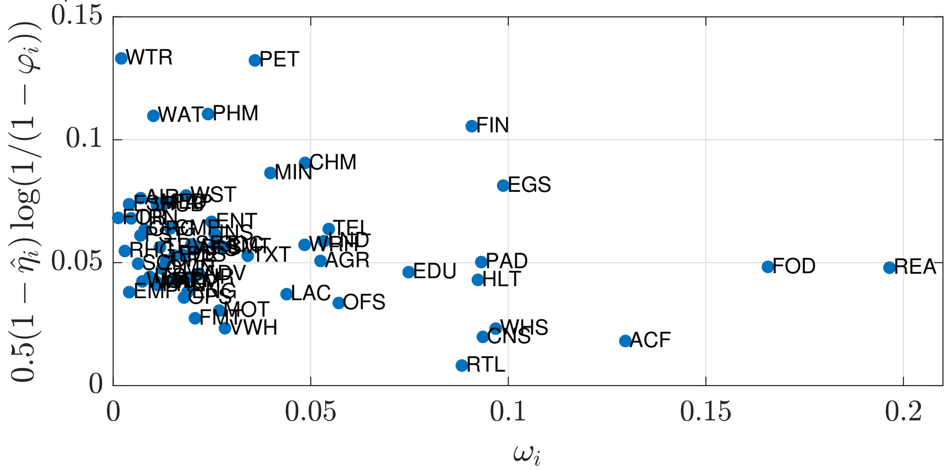


Figure 16 decomposes the gap in GDP between our baseline model and the fixed returns-to-scale economy in its sectoral components. It reports the two terms in (26) that captures a sector's importance: 1) its Domar weight ω_i and 2) the flexibility of its sectoral productivity $\frac{1}{2}(1 - \hat{\eta}_i) \log \frac{1}{1-\varphi_i}$. We see that the "Water transport" sector is the most flexible one. However, since its Domar weight is only 0.0021, its importance for the economy is small. High-Domar-weights sectors like "Finance", "Real estate", and "Electricity and gas" that are also flexible are where the endogenous returns-to-scale mechanism has the most impact on GDP.

B.5 Sensitivity analysis for Section 7.3

In Section 7.3, we experiment with removing wedges that are correlated with firm productivity. As we discuss there, removing these wedges leads to higher productivity dispersion. For some sectors, removing wedges would imply that $\varphi_i = \frac{\sigma_i^2}{2\gamma_i} > 1$, which is not allowed by our model. For these sectors, we set $\varphi_i = 0.99$. In this online supplement, we explore how sensitive our results are to this threshold. Table 4 shows log GDP gains due to removal of sales wedges if we set $\varphi_{max} = 0.985$, 0.99 (main text), and 0.995. We see that the GDP gains become larger as φ_{max} increases. In the model, having sectors with $\varphi_i \to 1$ is particularly valuable because they feature a larger mass of

Figure 16: Domar weights, ω_i, and productivity gain due to endogenous returns to scale, $\frac{1}{2}\left(1-\hat{\eta}_i\right)\log\frac{1}{1-\varphi_i}$, across sectors



firms with very high productivity draws operating at nearly constant returns to scale, which makes these sectors especially productive.

Table 4: Log GDP change after removal of sales wedges: Sensitivity analysis

	$\varphi_{max} = 0.985$	$\varphi_{max} = 0.99$	$\varphi_{max} = 0.995$
Baseline economy	160%	167%	177%
Dispersed RTS	134%	138%	142%
Fixed RTS	69%	70%	70%

Notes: Increases in log GDP due to removal of sales wedges in the baseline economy, and in the economies with fixed and dispersed returns to scale, for three values of maximum effective productivity dispersion φ.

C Proofs

C.1 Sectoral Domar weights

Multiplying the resource constraint for good i, given by (9), by P_i we get

$$P_iQ_i = P_iC_i + \sum_j P_i \int_0^{M_i} X_{ji,l}dl.$$

From the problem of the household we know that $P_i C_i = \beta_i \bar{P} Y$. It follows that

$$\frac{P_i Q_i}{\bar{P} Y} = \beta_i + \sum_j \frac{P_i}{\bar{P} Y} \int_0^{M_i} X_{ji,l} dl,$$

where we have divided by nominal GDP $\bar{P} Y$. Next, from the problem of firm l in sector j we know that

$$P_i X_{ji,l} = \alpha_{ji} \eta_{jl} P_j Q_{jl}.$$

Combining with the previous expression yields

$$\frac{P_i Q_i}{\bar{P} Y} = \beta_i + \sum_j \int_0^{M_i} \alpha_{ji} \eta_{jl} \frac{P_j Q_{jl}}{\bar{P} Y} dl,$$

or

$$\omega_i = \beta_i + \sum_j \alpha_{ji} \omega_j \hat{\eta}_j.$$

Solving this linear system leads to (18).

C.2 Proof of Lemma 1

Lemma 1. *The firm's marginal cost of production λ_{il} is given by*

$$\lambda_{il} = \frac{1}{e^{\varepsilon_{il}} A_i (\eta_{il})} H_i^{\eta_{il}} \Pi_{il}^{1-\eta_{il}}, \tag{3}$$

where $H_i := W^{1-\sum_{j=1}^{N} \alpha_{ij}} \prod_{j=1}^{N} P_j^{\alpha_{ij}}$ is the price of the variable input bundle in sector i, and Π_{il} is profits,

$$\Pi_{il} = (1 - \eta_{il}) \lambda_{il} Q_{il}. \tag{4}$$

Proof. We tackle problem (2) through its cost minimization dual:

$$\min_{\eta_{il}, L_{il;}, X_{ij,l}} W L_{il} + \sum_{j=1}^{N} P_j X_{ij,l}, \text{ subject to } F_i (L_{il}, X_{il}, \eta_{il}) \geq Q_{il}. \tag{49}$$

The Lagrangian is

$$\mathcal{L} = WL_{il} + \sum_{j=1}^{N} P_j X_{ij,l} - \lambda_{il} \left(e^{\varepsilon_{il}} A_i(\eta_{il}) \zeta(\eta_{il}) \left(L_{il}^{1-\sum_{j=1}^{N}\alpha_{ij}} \prod_{j=1}^{N} X_{ij,l}^{\alpha_{ij}} \right)^{\eta_{il}} - Q_{il} \right),$$

and the first-order conditions with respect to L_{il} and $X_{ij,l}$ are

$$\eta_{il} \left(1 - \sum_{j=1}^{N} \alpha_{ij}\right) \lambda_{il} Q_{il} = WL_{il}, \tag{50}$$

$$\eta_{il}\alpha_{ij}\lambda_{il}Q_{il} = P_j X_{ij,l}. \tag{51}$$

Plugging back into the constraint, we find

$$\lambda_{il} = \frac{1}{\left(e^{\varepsilon_{il}} A_i(\eta_{il})\right)^{\frac{1}{\eta_{il}}}} H_i\left(\left(1-\eta_{il}\right) Q_{il}\right)^{\frac{1-\eta_{il}}{\eta_{il}}}. \tag{52}$$

Using the definition of Π_{il} from (4) yields the result.

Note also that the envelope theorem implies that λ_{il} is the marginal production cost of the firm. Notice that λ_{il} is increasing in Q_{il} for $\eta_{il} < 1$. As usual, we can then write the profit maximization problem of the firm as

$$\max_{Q_{il}} P_i Q_{il} - \int_0^{Q_{il}} \lambda_{il}(x)\, dx,$$

where the notation makes clear the dependence of $\lambda_{il}(Q_{il})$ on the size of the firm. This problem's first-order condition implies that $P_i = \lambda_{il}(Q_{il})$, so that the firm sets Q_{il} to equalize its marginal cost to the price of its good. □

C.3 Proof of Lemma 2

Lemma 2. *At an interior solution, the firm chooses its returns to scale $\eta_{il} \in (0,1)$ according to*

$$\frac{da_i(\eta_{il})}{d\eta_{il}} = \log H_i - \log \Pi_{il}, \tag{5}$$

where $a_i(\eta_{il}) := \log A_i(\eta_{il})$.

.

Proof. The first-order condition for η_{il} in the cost-minimization problem (49) is

$$\frac{dA_i(\eta_{il})}{d\eta_{il}}\zeta(\eta_{il})\left(L_{il}^{1-\sum_{j=1}^N\alpha_{ij}}\prod_{j=1}^N X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}+A_i(\eta_{il})\frac{d\zeta(\eta_{il})}{d\eta_{il}}\left(L_{il}^{1-\sum_{j=1}^N\alpha_{ij}}\prod_{j=1}^N X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}} \tag{53}$$
$$+A_i(\eta_{il})\zeta(\eta_{il})\frac{d}{d\eta_{il}}\left(L_{il}^{1-\sum_{j=1}^N\alpha_{ij}}\prod_{j=1}^N X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}=0.$$

Note that we do not include Lagrange multipliers for the constraints $0\le\eta_{il}\le 1$ since we focus on interior solutions. Dividing by Q_{il} yields

$$\frac{d\log A_i(\eta_{il})}{d\eta_{il}}+\frac{d\log\zeta(\eta_{il})}{d\eta_{il}}+\frac{d}{d\eta_{il}}\log\left(L_{il}^{1-\sum_{j=1}^N\alpha_{ij}}\prod_{j=1}^N X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}=0.$$

Combining this with (50) and (51) yields (5). □

C.4 Proof of Lemma 3

Lemma 3. *At an interior solution, the returns-to-scale parameter η_{il} satisfies*[54]

$$\frac{d\eta_{il}}{d\varepsilon_{il}}=\frac{d\eta_{il}}{d\log P_i}=-\left[(1-\eta_{il})\frac{d^2a_i}{d\eta_{il}^2}\right]^{-1}>0,\quad \textit{and}\quad \frac{d\eta_{il}}{d\log H_i}=\left[(1-\eta_{il})\frac{d^2a_i}{d\eta_{il}^2}\right]^{-1}<0.$$

Proof. We can combine (3) with the firm's optimality condition $\lambda_i=P_i$ to write

$$\log\Pi_{il}=\frac{1}{1-\eta_{il}}\left(\log P_i+\varepsilon_{il}+a_i(\eta_{il})-\eta_{il}\log H_i\right). \tag{54}$$

Together with (5), we can write the first-order condition with respect to η_{il} as

$$\underbrace{\log H_i-\log P_i-\varepsilon_{il}}_{K}=(1-\eta_{il})\frac{d\log A_i(\eta_{il})}{d\eta_{il}}+\log A_i(\eta_{il}), \tag{55}$$

where we use K as a temporary variable to denote the left-hand side of (55). Full differentiation

[54]When increasing P_i, we keep the price of the variable input bundle constant to distinguish the two channels that affect η_{il}.

yields

$$1 = -\frac{d\eta_{il}}{dK}\frac{d\log A_i\left(\eta_{il}\right)}{d\eta_{il}} + \left(1-\eta_{il}\right)\frac{d^2\log A_i\left(\eta_{il}\right)}{d\eta_{il}^2}\frac{d\eta_{il}}{dK} + \frac{d\log A_i\left(\eta_{il}\right)}{d\eta_{il}}\frac{d\eta_{il}}{dK}.$$

Simplifying we find

$$\frac{d\eta_{il}}{dK} = \frac{1}{\left(1-\eta_{il}\right)\frac{d^2\log A_i(\eta_{il})}{d\eta_{il}^2}},$$

and the result follows. □

C.5 Proof of Lemma 4

Lemma 4. *At an interior solution, the elasticity of output Q_{il} with respect to productivity ε_{il} is given by*

$$\frac{d\log Q_{il}}{d\varepsilon_{il}} = \underbrace{\frac{1}{1-\eta_{il}}}_{\text{Fixed } \eta \text{ effect}} + \underbrace{\frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\varepsilon_{il}}}_{\text{Flexible } \eta \text{ effect}} > 0.$$

In addition, the elasticities of output Q_{il} with respect to prices are given by

$$\frac{d\log Q_{il}}{d\log P_i} = \underbrace{\frac{\eta_{il}}{1-\eta_{il}}}_{\text{Fixed } \eta \text{ effect}} + \underbrace{\frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log P_i}}_{\text{Flexible } \eta \text{ effect}} > 0, \quad \text{and} \quad \frac{d\log Q_{il}}{d\log H_i} = \underbrace{-\frac{\eta_{il}}{1-\eta_{il}}}_{\text{Fixed } \eta \text{ effect}} + \underbrace{\frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log H_i}}_{\text{Flexible } \eta \text{ effect}} < 0.$$

Proof. Profit maximization implies that the firm's marginal cost of production λ_i is equal to the price P_i, and so we can invert (3) and (4) to write

$$Q_{il} = \frac{1}{1-\eta_{il}}\left(e^{\varepsilon_{il}}A_i\left(\eta_{il}\right)\right)^{\frac{1}{1-\eta_{il}}}\left(\frac{P_i}{H_i}\right)^{\frac{\eta_{il}}{1-\eta_{il}}},$$

or, in log form, as

$$\log Q_{il} = -\log\left(1-\eta_{il}\right) + \frac{1}{1-\eta_{il}}\varepsilon_{il} + \frac{1}{1-\eta_{il}}a_i\left(\eta_{il}\right) + \frac{\eta_{il}}{1-\eta_{il}}\left(\log P_i - \log H_i\right). \tag{56}$$

Without endogenous returns to scale, it is immediate that

$$\frac{\partial\log Q_{il}}{\partial\varepsilon_{il}} = \frac{1}{1-\eta_{il}} \quad \text{and} \quad \frac{\partial\log Q_{il}}{\partial\log P_i} = -\frac{\partial\log Q_{il}}{\partial\log H_i} = \frac{\eta_{il}}{1-\eta_{il}}.$$

With endogenous returns to scale, we can combine (5) and (54) to find

$$-\left(1-\eta_{il}\right)\frac{da_i\left(\eta_{il}\right)}{d\eta_{il}}=\log P_i+\varepsilon_{il}+a_i\left(\eta_{il}\right)-\log H_i. \tag{57}$$

Combining (56) and (57), we get

$$\log Q_{il}=-\log\left(1-\eta_{il}\right)-\frac{da_i\left(\eta_{il}\right)}{d\eta_{il}}-\left(\log P_i-\log H_i\right). \tag{58}$$

Differentiating with respect to ε_{il}, we find

$$\frac{d\log Q_{il}}{d\varepsilon_{il}}=\frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\varepsilon_{il}}-\frac{d^2a_i\left(\eta_{il}\right)}{d\eta_{il}^2}\frac{d\eta_{il}}{d\varepsilon_{il}}.$$

Combining with Lemma 3 yields the result. The derivatives with respect to $\log P_i$ and $\log H_i$ can be computed in a similar way. The last part of the result follows from the signs of the derivatives in Lemma 3. □

C.6 Proof of Proposition 1

Proposition 1. *Suppose that Assumption 1 holds. Without endogenous returns to scale, the distribution of Q_{il} in sector i is log-normal. With endogenous returns to scale, the right tail of the distribution of Q_{il} behaves like a Pareto distribution with tail index $1/\varphi_i$, in the sense that*

$$\log\left(\mathbb{P}\left(Q_{il}>q\right)\right)\sim-\frac{1}{\varphi_i}\log q,\ \ \textit{as } q\to\infty.$$

Proof. Without endogenous returns to scale, the log of Q_{il} is given by (56). The only random term is ε_{il} and so Q_{il} is log-normal. We now turn to the case with endogenous returns to scale. Under Assumption 1, we can write (57) as

$$\frac{1}{1-\eta_{il}}=\frac{\varepsilon_{il}+B_i}{2\gamma_i},$$

where we define $B_i:=\log P_i-\log H_i$ as a temporary variable to simplify the notation. Combining with (58), we can write

$$\log Q_{il}=\log\left(\frac{\varepsilon_{il}+B_i}{2\gamma_i}\right)+\gamma_i\left(\frac{\varepsilon_{il}+B_i}{2\gamma_i}\right)^2-B_i.$$

We want to characterize the right tail of Q_{il}. Because of the logarithm, we need to be careful about eventual bounds on ε_{il}. We impose here that $\varepsilon_{il} \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ is normally distributed with a truncation such that $\varepsilon_{il} > -B_i$. We provide a full treatment of the model with truncated normal distribution in Supplement D.1. To simplify the notation, we drop the subscripts i and l from now on.

Step 1. We want to characterize the Complementary CDF (CCDF) $S_Q(q) = \mathbb{P}(Q > q)$ as $q \to \infty$. Let us define $g : (-B, \infty) \to \mathbb{R}$ as the function that maps ε to $\log Q$:

$$g(x) = \log\left(\frac{x+B}{2\gamma}\right) + \gamma\left(\frac{x+B}{2\gamma}\right)^2 - B.$$

One can show that g is a strictly increasing function. It is therefore invertible, and we can write

$$S_Q(q) = \mathbb{P}(\log Q > \log q) = \mathbb{P}(g(\varepsilon) > \log q) = \mathbb{P}\left(\varepsilon > g^{-1}(\log q)\right).$$

Given the properties of g, the right tail of Q corresponds to the right tail of ε.

Step 2. Let $y = g(x)$. We need to characterize the asymptotic behavior of $x = g^{-1}(y)$ as $y \to \infty$. Letting $X = \frac{x+B}{2\gamma}$, the equation $y = g(x)$ can be rewritten as

$$y + B = \log X + \gamma X^2.$$

As $y \to \infty$, it must be that $X \to \infty$. In this limit, the quadratic term γX^2 dominates $\log X$ and we can write[55]

$$y + B \sim \gamma X^2, \text{ as } y \to \infty.$$

This implies that $X \sim \sqrt{y/\gamma}$ for large y.

Now, we relate this to $x = g^{-1}(y)$. Since $x = 2\gamma X - B$, we have

$$g^{-1}(y) = x = 2\gamma X - B \sim 2\sqrt{\gamma y}.$$

since the constant B is negligible as $y \to \infty$. We will come back to this expression momentarily.

Step 3. The CCDF of the truncated normal ε is given by

$$S_\varepsilon(x) = \frac{1}{K_1} S_{\bar{\varepsilon}}(x)$$

[55] As usual, we write "$f(x) \sim g(x)$ as $x \to \infty$" if and only if $\lim_{x\to\infty} f(x)/g(x) = 1$.

where $\bar{\varepsilon}$ is the untruncated normal with the same mean and variance, and where K_1 is a constant. It is well-known that approximating the Mills ratio implies that

$$S_{\bar{\varepsilon}}(x) \sim \frac{\sigma}{(x-\mu)\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ as } x \to \infty.$$

We can therefore write

$$\log S_{\varepsilon}(x) \sim \log\left(\frac{1}{K_1}\right) - \frac{(x-\mu)^2}{2\sigma^2} + \log\left(\frac{\sigma}{(x-\mu)\sqrt{2\pi}}\right).$$

As $x \to \infty$, the quadratic term dominates the others and thus

$$\log S_{\varepsilon}(x) \sim -\frac{x^2}{2\sigma^2}, \text{ as } x \to \infty.$$

Step 4. We now combine the results. Let $x_q = g^{-1}(\log q)$. From Step 1, $S_Q(q) = S_{\varepsilon}(x_q)$. From Step 3, for large q, and consequently large x_q,

$$\log S_Q(q) \sim -\frac{x_q^2}{2\sigma^2}.$$

We can now substitute the asymptotic form for x_q from Step 2. Let $y = \log q$. As $q \to \infty$, $y \to \infty$ and

$$x_q = g^{-1}(\log q) \sim 2\sqrt{\gamma \log q}.$$

It is well-known that if $f \sim g$ then $f^r \sim g^r$ for r real. Therefore,

$$x_q^2 \sim 4\gamma \log q.$$

Since $\sim$ is transitive, we can substitute in the expression for $\log S_Q$ to find

$$\log S_Q(q) \sim -\frac{2\gamma}{\sigma^2} \log q, \text{ as } q \to \infty,$$

which is the result. $\square$

C.7 Proof of Lemma 5

Lemma 5. *The returns to scale η_{il} of firm l in sector i is given by*

$$\frac{1}{1-\eta_{il}} = \frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\varepsilon_{il}-\mu_i}{2\gamma_i}. \tag{12}$$

Furthermore, the moments of the firm-level returns-to-scale distribution are given by

$$\mathrm{E}_i\left[\frac{1}{1-\eta_{il}}\right] = \frac{1-\varphi_i}{1-\hat{\eta}_i}, \quad \mathrm{V}_i\left[\frac{1}{1-\eta_{il}}\right] = \frac{\varphi_i}{2\gamma_i}, \quad \text{and} \quad \mathrm{Cov}_i\left[\frac{1}{1-\eta_{il}}, \varepsilon_{il}\right] = \varphi_i > 0. \tag{13}$$

Proof. Given Assumption 1, we can write the returns to scale first-order condition (55) as

$$\log P_i - \log H_i + \varepsilon_{il} = \frac{2\gamma_i}{1-\eta_{il}},$$

Combining that expression with itself when $\varepsilon_{il} = \mu_i$ yields

$$\frac{1}{1-\eta_{il}} = \frac{1}{1-\eta_i(\mu_i)} + \frac{\varepsilon_{il}-\mu_i}{2\gamma_i},$$

and the result follows from combining with (64), derived below, and taking the moments. □

C.8 Proof of Proposition 2

Proposition 2. *The marginal cost of sector i is given by*

$$\lambda_i = \frac{1}{Z_i(\hat{\eta}_i)} W^{1-\hat{\eta}_i \sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N P_j^{\hat{\eta}_i \alpha_{ij}}, \tag{15}$$

where sectoral total factor productivity $Z_i(\hat{\eta}_i)$ is defined as

$$\log Z_i(\hat{\eta}_i) := \underbrace{\mu_i + a_i(\hat{\eta}_i) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i}}_{\text{Exogenous returns to scale}} + \underbrace{\frac{1}{2}(1-\hat{\eta}_i)\log\left(\frac{1}{1-\varphi_i}\right)}_{\text{Superstar effect}} - \underbrace{(1-\hat{\eta}_i)\log\kappa_i}_{\text{Entry cost}}. \tag{16}$$

Furthermore, the effective returns to scale $\hat{\eta}_i$ is given by

$$\frac{1}{1-\hat{\eta}_i} = \frac{1}{2\gamma_i(1-\varphi_i)}(\mu_i + \log P_i - \log H_i). \tag{17}$$

Proof. Since firms in a sector all face the same sales price, they have the same marginal cost through profit maximization. We therefore define the marginal cost λ_i of a sector i as the marginal cost of any firm in that sector, such that $\lambda_i := \lambda_{il}$ for any l.

Together with (54), the free-entry condition (8) imposes that

$$\int_{-\infty}^{\infty} \underbrace{\left(\lambda_i \frac{e^{\varepsilon_{il}} A_i(\eta_{il})}{H_i^{\eta_{il}}}\right)^{\frac{1}{1-\eta_{il}}}}_{\Pi_{il}} f_i(\varepsilon_{il})\, d\varepsilon_{il} = \kappa_i W, \tag{59}$$

where f_i is the probability density function of a normal distribution $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$. Multiplying the term inside the parentheses by one, we find

$$\int_{-\infty}^{\infty} \left(\frac{\lambda_i}{H_i^{\eta_{il}}} \frac{H_i^{(1-\eta_{il})\frac{\hat{\eta}_i}{1-\hat{\eta}_i}}}{H_i^{(1-\eta_{il})\frac{\hat{\eta}_i}{1-\hat{\eta}_i}}} \frac{\lambda_i^{\frac{1-\eta_{il}}{1-\hat{\eta}_i}}}{\lambda_i^{\frac{1-\eta_{il}}{1-\hat{\eta}_i}}} e^{\varepsilon_{il}} A_i(\eta_{il})\right)^{\frac{1}{1-\eta_{il}}} f_i(\varepsilon_{il})\, d\varepsilon_{il} = \kappa_i W,$$

which can be reorganized as

$$\lambda_i = \frac{1}{\tilde{Z}_i} (\kappa_i W)^{1-\hat{\eta}_i} \left(W^{1-\sum_j \alpha_{ij}} \prod_{j=1}^{N} P_j^{\alpha_{ij}}\right)^{\hat{\eta}_i}, \tag{60}$$

where $\tilde{Z}_i$ is defined as

$$\tilde{Z}_i := \left[\int_{-\infty}^{\infty} \left(\left(\frac{\lambda_i}{H_i}\right)^{\frac{\eta_{il}-\hat{\eta}_i}{1-\hat{\eta}_i}} e^{\varepsilon_{il}} A_i(\eta_{il})\right)^{\frac{1}{1-\eta_{il}}} f_i(\varepsilon_{il})\, d\varepsilon_{il}\right]^{1-\hat{\eta}_i}.$$

To simplify the notation, define $s_i := \log \lambda_i - \log H_i$. Using the definition of s_i and A_i, we can write

$$\tilde{Z}_i = \left[\int_{-\infty}^{\infty} \left(e^{s_i \frac{\eta_{il}-\hat{\eta}_i}{1-\hat{\eta}_i} + \varepsilon_{il} - \frac{\gamma_i}{1-\eta_{il}}}\right)^{\frac{1}{1-\eta_{il}}} f_i(\varepsilon_{il})\, d\varepsilon_{il}\right]^{1-\hat{\eta}_i}. \tag{61}$$

For an arbitrary set of firm-level returns to scale $\{\eta_{il}\}$ this integral cannot be computed analytically, but we can do so here, given the relationship between ε_{il} and η_{il} implied by the model. Using

Assumption 1, we can write the returns to scale first-order condition (55) as[56]

$$\underbrace{\log \lambda_i - \log H_i}_{:=s_i} + \varepsilon_{il} = \frac{2\gamma_i}{1-\eta_{il}} = -2a_i\left(\eta_{il}\right), \tag{62}$$

which implies that

$$\frac{1}{1-\eta_{il}} = \frac{s_i + \varepsilon_{il}}{2\gamma_i}, \text{ and } \frac{\eta_{il}}{1-\eta_{il}} = \frac{s_i + \varepsilon_{il} - 2\gamma_i}{2\gamma_i}.$$

Combining with $\tilde{Z}_i$, we find

$$\tilde{Z}_i = \left[\int_{-\infty}^{\infty} e^{\frac{(s_i+\varepsilon_{il})^2}{4\gamma_i} - \frac{s_i}{1-\hat{\eta}_i}} f_i\left(\varepsilon_{il}\right) d\varepsilon_{il}\right]^{1-\hat{\eta}_i}.$$

Given the structure of the normal distribution f_i, this integral can be computed when $2\gamma_i > \sigma_i^2$ and yields

$$\tilde{Z}_i = \left[\sqrt{\frac{2\gamma_i}{2\gamma_i - \sigma_i^2}} \exp\left(\frac{\left(s_i + \mu_i\right)^2}{2\left(2\gamma_i - \sigma_i^2\right)} - \frac{s_i}{1-\hat{\eta}_i}\right)\right]^{1-\hat{\eta}_i}.$$

We will rewrite this expression using $\hat{\eta}_i$. To do so, notice that we can write

$$\hat{\eta}_i = \frac{\int_l \eta_{il} P_i Q_{il} dl}{\int_l P_i Q_{il} dl} = 1 - \frac{\int_l \left(1-\eta_{il}\right) P_i Q_{il} dl}{\int_l P_i Q_{il} dl} = 1 - \frac{\int_l \Pi_{il} dl}{\int_l \frac{1}{1-\eta_{il}} \Pi_{il} dl}. \tag{63}$$

Using the profit expression (54), we can compute these integrals and find

$$1 - \hat{\eta}_i = 2\gamma_i \frac{1-\varphi_i}{s_i + \mu_i} = \left(1-\varphi_i\right)\left(1 - \eta_i\left(\mu_i\right)\right), \tag{64}$$

where $\eta_i\left(\mu_i\right)$ is the returns to scale chosen by the firm with $\varepsilon_{il} = \mu_i$ (computed from (62)). Notice that (64) implies (17) because s_i is given by (62).

Combining (64) with our expression for $\tilde{Z}_i$, we find

$$\tilde{Z}_i = \left[\sqrt{\frac{1}{1-\varphi_i}} \exp\left(\frac{1-\varphi_i}{1-\hat{\eta}_i} a_i\left(\hat{\eta}_i\right) + \frac{\mu_i}{1-\hat{\eta}_i}\right)\right]^{1-\hat{\eta}_i}.$$

[56]In equilibrium, the price charged by firms in sector i must be equal to their marginal costs, so that $\lambda_i = P_i$.

Taking the log yields

$$\log \tilde{Z}_i(\hat{\eta}_i) := \mu_i + a_i(\hat{\eta}_i) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i} - (1-\hat{\eta}_i)\log\left(\sqrt{1-\varphi_i}\right),$$

where we have used the definition of φ_i and Assumption 1. The quantity $\tilde{Z}_i$ corresponds to the total factor productivity of sector i if we treat the mass of firms in that sector as an independent factor. But it will be often convenient to lump that input together with labor. In that case, we can rewrite (60) as

$$\lambda_i = \frac{1}{Z_i(\hat{\eta}_i)} W^{1-\hat{\eta}_i \sum_{j=1}^{N}\alpha_{ij}} \left(\prod_{j=1}^{N} P_j^{\alpha_{ij}}\right)^{\hat{\eta}_i},$$

where $Z_i := \tilde{Z}_i(\hat{\eta}_i)/\kappa_i^{1-\hat{\eta}_i}$, which completes the proof. □

C.9 Proof of Proposition 3

Proposition 3. *The equilibrium price vector* $P = (P_1, \ldots, P_N)$ *satisfies*

$$\log \frac{P}{W} = -\mathcal{L}(\hat{\eta})\, z(\hat{\eta}), \tag{19}$$

where $z(\hat{\eta}) = (\log Z_1(\hat{\eta}_1), \ldots, \log Z_N(\hat{\eta}_N))$ *is the vector of log sectoral productivities* (16). *Furthermore, equilibrium log GDP* $y := \log Y$ *is given by*

$$y(\hat{\eta}) = \underbrace{[\omega(\hat{\eta})]^{\top} z(\hat{\eta})}_{\textit{Aggregate productivity}} + \underbrace{\log \bar{L}}_{\textit{Labor endowment}}. \tag{20}$$

Proof. Since in equilibrium prices must be equal to marginal costs, we can use (15) to write

$$\frac{P_i}{W} = \frac{1}{Z_i(\hat{\eta}_i)} \prod_{j=1}^{N}\left(\frac{P_j}{W}\right)^{\hat{\eta}_i \alpha_{ij}}.$$

Taking the log of this equation leads to

$$\log\frac{P_i}{W} = -\log Z_i(\hat{\eta}_i) + \hat{\eta}_i \sum_{j=1}^{N} \alpha_{ij} \log\frac{P_j}{W}.$$

In vector notation, this becomes $\log(P/W) = -z(\hat{\eta}) + \mathrm{diag}(\hat{\eta})\,\alpha \log(P/W)$. Solving it for

$\log(P/W)$ yields (19).

We now turn to the GDP equation. The budget constraint of the household is $\bar{P}Y = W\bar{L}$. Together with the definition of the price index, $\bar{P} = \prod_{i=1}^{N} P_i^{\beta_i} = 1$, we can therefore write

$$y = -\sum_{i=1}^{N} \beta_i \log \frac{P_i}{W} + \log \bar{L},$$

and the result follows from combining this expression with (18) and (19). □

C.10 Proof of Proposition 4

Proposition 3. *There exists a unique equilibrium, and it is efficient. Furthermore, the equilibrium vector of effective returns to scale $\hat{\eta}$ maximizes GDP $y(\hat{\eta})$.*

Proof. This proof proceeds in two steps. First, we write down the maximization problem of the social planner and show that its first-order conditions coincide with the equilibrium conditions. Since there exists at least one maximizer to the planner's problem, there is at least one solution to the planner's first-order conditions and so at least one efficient equilibrium exists. Second, we show that the equilibrium conditions imply that there can be at most one equilibrium.

Step 1. The planner maximizes

$$\max_{C,X,L,M,\eta} \sum_{i=1}^{N} \beta_i \log(C_i)$$

subject to the goods resource constraint

$$C_i + \sum_{j=1}^{N} M_j \int X_{ji}(\varepsilon) f_j(\varepsilon) d\varepsilon \leq M_i \int Q_i(\varepsilon) f_i(\varepsilon) d\varepsilon \quad \forall i \in \{1, ..., N\} \quad \text{(multiplier } \lambda_i\text{)},$$

and the labor resource constraint

$$\sum_{i=1}^{N} M_i \int L_i(\varepsilon) f_i(\varepsilon) d\varepsilon + \sum_{i=1}^{N} M_i \kappa_i \leq \bar{L} \quad \text{(multiplier } \mu\text{)}.$$

The first-order necessary conditions are as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial C_i}: &\quad \lambda_i = \frac{\beta_i}{C_i}, \\
\frac{\partial \mathcal{L}}{\partial L_i(\varepsilon)}: &\quad \lambda_i \frac{\partial Q_i(\varepsilon)}{\partial L_i(\varepsilon)} - \mu = 0, \\
\frac{\partial \mathcal{L}}{\partial X_{ij}(\varepsilon)}: &\quad \lambda_i \frac{\partial Q_i(\varepsilon)}{\partial X_{ij}(\varepsilon)} - \lambda_j = 0, \\
\frac{\partial \mathcal{L}}{\partial \eta_i(\varepsilon)}: &\quad \frac{\partial Q_i(\varepsilon)}{\partial \eta_i(\varepsilon)} = 0, \\
\frac{\partial \mathcal{L}}{\partial M_i}: &\quad \int \left[\lambda_i Q_i(\varepsilon) - \sum_{j=1}^{N} \lambda_j X_{ij}(\varepsilon) - \mu L_i(\varepsilon) \right] f_i(\varepsilon) d\varepsilon = \mu \kappa_i.
\end{aligned}
$$

Now, we demonstrate that the competitive equilibrium allocation satisfies the planner's optimality conditions. To do this, we identify the planner's shadow prices with the equilibrium market prices. Set $\mu = W$. Consequently, the planner's shadow price for good i, λ_i, corresponds to the market price P_i. The first condition corresponds to the household's optimality condition (Section 2.4). The second and third optimality conditions correspond to the standard firm equilibrium optimality conditions (50) and (51). The planner's fourth optimality condition coincides with the firm equilibrium condition (53). Finally, the last optimality condition of the planner coincide with the free entry condition (8). Since the resource constraints are the same in the planner's problem and the equilibrium definition, we have shown that the planner's first-order conditions coincides with the equilibrium conditions. Since the planner's constraint set is closed and bounded, and that the objective function is continuous, the Extreme Value Theorem implies that there exists a maximizer to the planner's problem. This maximizer must satisfy the first-order necessary conditions. Therefore, there exists an equilibrium and that equilibrium is efficient.

Step 2. We now show that there can be at most one equilibrium. The equilibrium of the model boils down to equations (17) and (19). Indeed, if we let $p := \log(P/W)$ we can write these equations as

$$p = -\mathcal{L}(\hat{\eta}) z(\hat{\eta}), \tag{65}$$

where

$$z_i(\hat{\eta}) := \mu_i + a_i(\hat{\eta}_i) + \frac{\sigma_i^2}{2} \frac{1}{1-\hat{\eta}_i} + (1-\hat{\eta}_i) \log\left(\frac{1}{\sqrt{1-\varphi_i}}\right) - (1-\hat{\eta}_i) \log \kappa_i, \tag{66}$$

and

$$\frac{1}{1-\hat{\eta}_i} = \frac{1}{2\gamma_i\left(1-\varphi_i\right)}\left(\mu_i + p_i - \sum_{j=1}^{N}\alpha_{ij}p_j\right). \tag{67}$$

There is a unique equilibrium if there are unique vectors $\hat{\eta}$ and p that solve these equations. We can combine these equations into a single one. Let us introduce the variable $v := (I-\alpha)\,p$ and a constant $C_i = 2\gamma_i\left(1-\varphi_i\right) > 0$. We can then rewrite (67) as

$$1-\hat{\eta}_i = \frac{C_i}{\mu_i + v_i} \Leftrightarrow \hat{\eta}_i = 1 - \frac{C_i}{\mu_i + v_i}.$$

We are interested in equilibrium of the firm $0 < \hat{\eta}_i < 1$ for all i.[57] This implies that we can restrict the relevant domain of v to be

$$\mu_i + v_i > C_i.$$

Using that notation, we can simplify the equation (66) for z_i as

$$z_i = \frac{\mu_i - v_i}{2} + \frac{C_i}{\mu_i + v_i}\log\left(K_i/\kappa_i\right),$$

where $K_i := 1/\sqrt{1-\varphi_i}$. Next, we can premultiply (65) by $\mathcal{L}\left(\hat{\eta}\right)^{-1} = \left(I - \operatorname{diag}\left(\hat{\eta}\right)\alpha\right)$ to find

$$\left(I - \operatorname{diag}\left(\hat{\eta}\right)\alpha\right)\left(I-\alpha\right)^{-1}v = -z$$

or

$$\left(I + \operatorname{diag}\left(1-\hat{\eta}\right)\alpha\left(I-\alpha\right)^{-1}\right)v = -z.$$

Substituting the expression for z and $1-\hat{\eta}$, we find

$$F_i\left(v\right) := \frac{1}{2}\left(\mu_i + v_i\right)^2 + C_i\left(\left(\alpha\left(I-\alpha\right)^{-1}v\right)_i + \log\left(K_i/\kappa_i\right)\right) = 0.$$

There is a unique equilibrium if there is a unique solution v to the equation $F\left(v\right) = 0$. Recall that $p = \left(I-\alpha\right)^{-1}v$. Then

$$\hat{F}_i\left(p\right) := F_i\left(v\left(p\right)\right) = \frac{1}{2}\left(\mu_i + \left(\left(I-\alpha\right)p\right)_i\right)^2 + C_i\left(\left(\alpha p\right)_i + \log\left(K_i/\kappa_i\right)\right).$$

[57]It is straightforward to write sufficient conditions on the parameters so that the equilibrium is of that form. In particular, large μ lead to higher equilibrium $\hat{\eta}$.

The Jacobian of $\hat{F}$ is

$$M_{ik}(p) = (\mu_i + ((I-\alpha)p)_i)(I-\alpha)_{ik} + C_i\alpha_{ik}.$$

In matrix form,

$$M(p) = \text{diag}(\mu + v(p))(I-\alpha) + \text{diag}(C)\alpha.$$

The diagonal elements of M are

$$M_{ii} = (\mu_i + v_i)(1-\alpha_{ii}) + C_i\alpha_{ii} > 0,$$

which is positive given our domain restriction that $\mu_i + v_i > C_i$. For off-diagonal terms $i \neq k$,

$$M_{ik} = -(\mu_i + v_i)\alpha_{ik} + C_i\alpha_{ik} = \alpha_{ik}(C_i - (\mu_i + v_i)) < 0,$$

such that M is a Z-matrix. Further notice that

$$\sum_{k\neq i} |M_{ik}| = ((\mu_i + v_i) - C_i)\sum_{k\neq i}\alpha_{ik}.$$

For M to be strictly diagonally dominant, it must be that

$$(\mu_i + v_i)(1-\alpha_{ii}) + C_i\alpha_{ii} > ((\mu_i + v_i) - C_i)\sum_{k\neq i}\alpha_{ik},$$

which we can reorganize as

$$\mu_i + v_i > -C_i\frac{\sum_k \alpha_{ik}}{1-\sum_k \alpha_{ik}}.$$

This condition is true since $C_i > 0$ and $\mu_i + v_i > C_i$. Therefore, M is diagonally dominant. It follows that $M(p)$ is a non-singular M-matrix for every p. Since nonsingular M-matrices are a subset of P-matrices, $M(p)$ is also a P-matrix for every p. By the Gale and Nikaido (1965) theorem, $\hat{F}(p)$ is therefore injective and can have at most one solution $\hat{F}(p) = 0$. There is therefore a unique p that solves our original system of equations. From the vector p, it is straightforward to recover all other equilibrium quantities in a unique fashion. There is therefore a unique equilibrium and it is efficient. □

C.11 Proof of Lemma 6

Lemma 6. *An increase in average productivity μ_j increases returns to scale in all other sectors, such that*

$$\frac{d\hat{\eta}_i}{d\mu_j} = \Psi_i^{-1}\mathcal{K}_{ij} \geq 0. \tag{22}$$

Furthermore, the impact of productivity dispersion σ_j^2 on $\hat{\eta}_i$ is given by

$$\frac{d\hat{\eta}_i}{d\sigma_j^2} = \Psi_i^{-1}\left(\mathcal{K}_{ij}\frac{\partial z_j}{\partial \sigma_j^2} - \mathbb{1}\left(i=j\right)\frac{\partial^2 z_i}{\partial \sigma_i^2 \partial \hat{\eta}_i}\right), \tag{23}$$

where

$$\frac{\partial z_j}{\partial \sigma_j^2} = \frac{1}{2\left(1-\hat{\eta}_j\right)} + \frac{1-\hat{\eta}_j}{4\gamma_j\left(1-\varphi_j\right)} > 0, \text{ and } \frac{\partial}{\partial \sigma_i^2}\left(\frac{\partial z_i}{\partial \hat{\eta}_i}\right) = \frac{1}{2\left(1-\hat{\eta}_i\right)^2} - \frac{1}{4\gamma_i\left(1-\varphi_i\right)}.$$

In particular, $d\hat{\eta}_i/d\sigma_j^2 \geq 0$ for $i \neq j$.

Proof. This proof proceeds as follows. First we derive the first-order conditions of the social planner. Second, we write down the derivative of the first-order conditions with respect to $\hat{\eta}_i$. Third, we use this expression together with the implicit function theorem to derive the impact of μ_j and σ_j^2 on $\hat{\eta}_i$.

First step. Let us first compute the first-order conditions of the planner's problem. Differentiating (20) with respect to $\hat{\eta}_i$ and setting that expression to zero implies that

$$\frac{dy}{d\hat{\eta}_i} = \beta^\top \frac{d\mathcal{L}}{d\hat{\eta}_i} z\left(\hat{\eta}\right) + \left[\omega\left(\hat{\eta}\right)\right]^\top \frac{dz\left(\hat{\eta}\right)}{d\hat{\eta}_i} = 0.$$

Computing the derivative of $z\left(\hat{\eta}\right)$, we find

$$\left(\frac{dz\left(\hat{\eta}\right)}{d\hat{\eta}_i}\right)_j = \begin{cases} \frac{da_i(\hat{\eta}_i)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{(1-\hat{\eta}_i)^2} + \frac{1}{2}\log\left(1-\varphi_i\right) + \log\kappa_i & \text{if } j=i \\ 0 & \text{if } j \neq i \end{cases}$$

Next, the derivative of the Leontief inverse yields

$$\begin{aligned} \frac{d\mathcal{L}}{d\hat{\eta}_i} &= \frac{d\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}}{d\hat{\eta}_i} = -\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}\left[\frac{d\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)}{d\hat{\eta}_i}\right]\left(1-\operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}. \\ &= \mathcal{L}\left[1_i 1_i^\top \alpha\right]\mathcal{L} = \mathcal{L}_{\cdot i}\alpha_i^\top \mathcal{L}. \end{aligned}$$

Putting the pieces together, we have

$$\omega_i(\hat{\eta})\,\alpha_i^\top \mathcal{L}(\hat{\eta})\,z(\hat{\eta}) + \omega_i(\hat{\eta})\left[\frac{da_i(\hat{\eta}_i)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{(1-\hat{\eta}_i)^2} + \frac{1}{2}\log(1-\varphi_i) + \log\kappa_i\right] = 0.$$

Since Domar weights are positive, we can write that condition as

$$\mathcal{F}_i := \alpha_i^\top \mathcal{L}(\hat{\eta})\,z(\hat{\eta}) + \frac{da_i(\hat{\eta}_i)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{(1-\hat{\eta}_i)^2} + \frac{1}{2}\log(1-\varphi_i) + \log\kappa_i = 0, \tag{68}$$

where we have defined $\mathcal{F}_i$.

Second step. The implicit function theorem states that

$$\frac{d\hat{\eta}}{d\mu} = -\left[\frac{\partial \mathcal{F}}{\partial \hat{\eta}}\right]^{-1}\left[\frac{\partial \mathcal{F}}{\partial \mu}\right].$$

First, let us compute the Jacobian matrix $\partial\mathcal{F}/\partial\hat{\eta}$. Consider an off-diagonal element $k \neq i$

$$\begin{aligned}\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_k} &= \frac{\partial}{\partial \hat{\eta}_k}\left(\alpha_i^\top \mathcal{L}(\hat{\eta})\,z(\hat{\eta})\right) = \alpha_i^\top\left(\frac{\partial \mathcal{L}}{\partial \hat{\eta}_k}\right)z + \alpha_i^\top \mathcal{L}\left(\frac{\partial z}{\partial \hat{\eta}_k}\right) \\ &= \alpha_i^\top \mathcal{L}\mathbf{1}_k\mathbf{1}_k^\top \alpha\mathcal{L}z + \alpha_i^\top \mathcal{L}\mathbf{1}_k\frac{\partial z_k}{\partial \hat{\eta}_k}.\end{aligned}$$

Factoring this expression gives

$$\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_k} = \left(\alpha_i^\top \mathcal{L}_{\cdot k}\right)\left[\alpha_k^\top \mathcal{L}z + \frac{\partial z_k}{\partial \hat{\eta}_k}\right] = 0,$$

where the last equality follows since the term in bracket is the first-order condition of the planner.

For a diagonal element,

$$\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_i} = \frac{\partial}{\partial \hat{\eta}_i}\left(\alpha_i^\top \mathcal{L}z + \frac{\partial z_i}{\partial \hat{\eta}_i}\right).$$

Through the logic above, the first term is 0, so we need only focus on the second part

$$\frac{\partial \mathcal{F}_i}{\partial \hat{\eta}_i} = \frac{\partial}{\partial \hat{\eta}_i}\frac{\partial z_i}{\partial \hat{\eta}_i} = \frac{d^2 a_i}{d\hat{\eta}_i^2} + \frac{\sigma_i^2}{(1-\hat{\eta}_i)^3} = (1-\varphi_i)\,\frac{d^2 a_i}{d\hat{\eta}_i^2}. \tag{69}$$

Third step. Next, we can compute the element (i, j) of the matrix $\frac{\partial \mathcal{F}}{\partial \mu}$, which is $\frac{\partial \mathcal{F}_i}{\partial \mu_j}$. The FOC for sector i is $\mathcal{F}_i = \alpha_i^\top \mathcal{L}z + \frac{\partial z_i}{\partial \hat{\eta}_i}$. The parameter μ_j only enters through the vector z, specifically through its j-th element. Therefore,

$$\frac{\partial z}{\partial \mu_j} = \mathbf{1}_j.$$

Thus,

$$\frac{\partial \mathcal{F}_i}{\partial \mu_j} = \frac{\partial}{\partial \mu_j}\left(\alpha_i^\top \mathcal{L} z\right) = \alpha_i^\top \mathcal{L}\left(\frac{\partial z}{\partial \mu_j}\right) = \alpha_i^\top \mathcal{L}\mathbf{1}_j,$$

which is simply the (i,j)-th element of the matrix $\alpha\mathcal{L}$. Putting the pieces together,

$$\frac{d\hat{\eta}}{d\mu} = -\left(\frac{\partial \mathcal{F}}{\partial \hat{\eta}}\right)^{-1}(\alpha\mathcal{L}) = -\frac{(\alpha\mathcal{L})_{ij}}{\frac{d^2 a_i}{d\hat{\eta}_i^2} + \frac{\sigma_i^2}{(1-\hat{\eta}_i)^3}} = -\left((1-\varphi_i)\frac{d^2 a_i}{d\hat{\eta}_i^2}\right)^{-1}(\alpha\mathcal{L})_{ij}\,.$$

Fourth step. We now turn to the impact of σ_j^2. We use the implicit function theorem once more. Note that

$$\frac{\partial \mathcal{F}_i}{\partial \sigma_j^2} = \alpha_i^\top \mathcal{L}\frac{\partial z}{\partial \sigma_j^2} + \frac{\partial}{\partial \sigma_j^2}\left(\frac{\partial z_i}{\partial \hat{\eta}_i}\right).$$

The vector $\partial z/\partial \sigma_j^2$ is zero everywhere except for it j-th element

$$\frac{\partial z_j}{\partial \sigma_j^2} = \frac{\partial}{\partial \sigma_j^2}\left(\frac{\sigma_j^2}{2(1-\hat{\eta}_j)} - \frac{1-\hat{\eta}_j}{2}\log(1-\varphi_j)\right) = \frac{1}{2(1-\hat{\eta}_j)} + \frac{1-\hat{\eta}_j}{4\gamma_j(1-\varphi_j)} > 0.$$

Similarly, $\frac{\partial}{\partial \sigma_j^2}\left(\frac{\partial z_i}{\partial \hat{\eta}_i}\right)$ is zero whenever $i \neq j$. We can compute

$$\frac{\partial}{\partial \sigma_j^2}\left(\frac{\partial z_j}{\partial \hat{\eta}_j}\right) = \frac{\partial}{\partial \sigma_j^2}\left(\frac{\sigma_j^2}{2(1-\hat{\eta}_j)^2} + \frac{1}{2}\log(1-\varphi_j)\right) = \frac{1}{2(1-\hat{\eta}_j)^2} - \frac{1}{4\gamma_j(1-\varphi_j)}.$$

Putting the pieces together, we find the result. □

C.12 Proof of Proposition 5

Proposition 5. *The difference in log GDP between the baseline model and the fixed returns-to-scale economy is given by*

$$y - \tilde{y} = \sum_{i=1}^{N}\frac{1}{2}\omega_i(1-\hat{\eta}_i)\log\left(\frac{1}{1-\varphi_i}\right) > 0.26$$

Proof. We first compute GDP in the fixed returns-to-scale economy (denoted by $\tilde{\cdot}$), in which all firms in sector i have the same returns to scale $\eta_{il} = \hat{\eta}_i$. The free-entry condition is $\mathrm{E}\left[\tilde{\Pi}_{il}\right] = \kappa_i \tilde{W}$.

Using the expression for profit (54), this condition becomes

$$\int_{-\infty}^{\infty} \exp\left(\frac{1}{1-\hat{\eta}_i}\left(\log \tilde{P}_i + \varepsilon_{il} + a_i\left(\hat{\eta}_i\right) - \hat{\eta}_i \log H_i\right)\right) f\left(\varepsilon_{il}\right) d\varepsilon_{il} = \kappa_i \tilde{W}.$$

Solving the integral and following the same aggregation steps as in the baseline model (Proposition 2), but without the endogenous choice of returns to scale, this condition yields a sectoral productivity of

$$\tilde{z}_i = \mu_i + a_i\left(\hat{\eta}_i\right) + \frac{\sigma_i^2}{2\left(1-\hat{\eta}_i\right)} - \left(1-\hat{\eta}_i\right) \log \kappa_i.$$

Because $\hat{\eta}_{il}$ is fixed, the term related to the choice of scale and the resulting amplified selection (i.e., the fourth term on the right-hand side of (16)) is absent. Since the sectoral production function and cost shares are still governed by $\hat{\eta}_i$, the pricing equation is analogous to the baseline model: $\log\left(\tilde{P}/\tilde{W}\right) = -\mathcal{L}\left(\hat{\eta}\right)\tilde{z}$. Log GDP is therefore given by:

$$\tilde{y} = \left[\omega\left(\hat{\eta}\right)\right]^{\top} \tilde{z}\left(\hat{\eta}\right) + \log \bar{L}.$$

Note that the Domar weights $\omega\left(\hat{\eta}\right)$ are identical to the baseline model because the sectoral input shares are the same in both economies.

Recall from 16 and 20 that in the baseline model

$$y = \left[\omega\left(\hat{\eta}\right)\right]^{\top} z\left(\hat{\eta}\right) + \log \bar{L},$$

where

$$z_i = \mu_i + a_i\left(\hat{\eta}_i\right) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i} + \frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\varphi_i}\right) - \left(1-\hat{\eta}_i\right)\log\kappa_i.$$

As a result,

$$y - \tilde{y} = \left[\omega\left(\hat{\eta}\right)\right]^{\top}\left(z\left(\hat{\eta}\right) - \tilde{z}\left(\hat{\eta}\right)\right). \qquad (70)$$

The difference in the sectoral productivity vectors, $z - \tilde{z}$, is a vector where the i-th element is

$$z_i\left(\hat{\eta}_i\right) - \tilde{z}_i\left(\hat{\eta}_i\right) = \frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\varphi_i}\right).$$

Substituting in (70) yields (26). The inequality $y - \tilde{y} > 0$ holds since $0 < \varphi_i < 1$ for all i. □

C.13 Proof of Proposition 7

Proposition 7. *The response of log GDP y to a shock $\Delta\mu_i$ is given by*

$$\Delta y = \omega_i \Delta\mu_i + \frac{1}{2}\frac{d\omega_i}{d\mu_i}\left(\Delta\mu_i\right)^2 + o\left(\left(\Delta\mu_i\right)^2\right). \quad (30)$$

Furthermore, the second-order term is non-negative,

$$\frac{d\omega_i}{d\mu_i} = \left(-\sum_{k=1}^{N}\mathcal{K}_{ki}\omega_k\frac{d\hat{\eta}_k}{d\mu_i}\right) \geq 0.$$

Proof. The second-order expansion of y with respect to productivity shocks is

$$\Delta y = \sum_{i=1}^{N}\frac{dy}{d\mu_i}\Delta\mu_i + \frac{1}{2}\sum_{i,j}^{N}\frac{d^2y}{d\mu_i d\mu_j}\Delta\mu_i\Delta\mu_j + o\left(\Delta^2\mu\right).$$

By Proposition 6, $dy/d\mu_i = \omega_i$ which yields (30) when $\Delta\mu_j = 0$ for all $j \neq i$. Next, Corollary 7 implies that

$$\frac{d\omega_i}{d\mu_i} = -\sum_{k=1}^{N}\mathcal{K}_{ki}\omega_k\frac{d\hat{\eta}_k}{d\mu_i} \geq 0,$$

where the inequality follows since $d\hat{\eta}_k/d\mu_i \geq 0$ from Corollary 6. □

C.14 Proof of Lemma 8

Lemma 8. *An increase in τ_j^S decreases the returns to scale in all downstream sectors:*

$$\frac{d\hat{\eta}_i}{d\tau_j^S} = -\frac{1}{1-\tau_j^S}\Psi_i^{-1}\mathcal{K}_{ij} \leq 0. \quad (32)$$

Proof. See proof of Proposition 10 in Supplement D.5. □

C.15 Proof of Proposition 8

Proposition 8. *In the presence of sales wedges, the impact of a parameter $\chi \in \left\{\mu_j, \sigma_j^2, \kappa_j, \gamma_j\right\}$ on GDP is given by*

$$\frac{dy}{d\chi} = \frac{\partial y}{\partial\chi} + \sum_{i=1}^{N}\frac{\partial y}{\partial\hat{\eta}_i}\frac{d\hat{\eta}_i}{d\chi},$$

6910where $\partial y/\partial\chi$ is given by Proposition 6, $d\hat{\eta}_i/\partial\chi$ is given by Corollaries 6 to 10, and $\partial y/\partial\hat{\eta}_i \geq 0$.

Proof. See proof of Proposition 12 in Supplement D.5. □

C.16 Proof of Corollary 1

Corollary 1. *The growth of effective returns to scale $\hat{\eta}$ is given by*

$$\frac{d\hat{\eta}}{dt} = \Psi^{-1}\mathcal{K}g_{\mu} > 0. \tag{34}$$

Furthermore, as $t \to \infty$, effective returns to scale $\hat{\eta}$ converges to 1.

Proof. The first equation follows directly from (22). Note that the right-hand side is strictly positive for $0 < \hat{\eta} < 1$ and converges to 0 as $\hat{\eta} \to 1$. The second result follows. □

C.17 Proof of Proposition 9

Proposition 9. *The growth rate of GDP is given by*

$$\frac{dy}{dt} = \frac{g_{\mu}}{1-\alpha} \times \left(1 - \frac{1}{\sqrt{\frac{1}{\gamma}\frac{1-\alpha}{\alpha}\frac{g_{\mu}}{1-\varphi}t + K}}\right) > 0, \tag{35}$$

where $K > 0$ is a time-invariant term given in the proof of the proposition.

Proof. The envelope theorem implies that

$$\frac{dy}{dt} = \left(1 - \hat{\eta}\alpha\right)^{-1} g_{\mu}. \tag{71}$$

Therefore, to characterize dy/dt, we need to solve for $\hat{\eta}\left(t\right)$. Equation (34) can be written as

$$\frac{d\hat{\eta}}{dt} = \frac{1}{2\gamma - \sigma^2}\frac{\alpha\left(1-\hat{\eta}\right)^3}{1-\hat{\eta}\alpha}g_{\mu}$$

and reorganized as

$$\frac{\alpha g_{\mu}}{\left(\gamma - \frac{\sigma^2}{2}\right)2}dt = \left(\frac{1-\alpha}{\left(1-\hat{\eta}\right)^3} + \frac{\alpha}{\left(1-\hat{\eta}\right)^2}\right)d\hat{\eta}.$$

Integrating on both sides yields

$$\frac{\alpha g_\mu}{\left(\gamma - \frac{\sigma^2}{2}\right)2}t + R = \frac{1-\alpha}{2\left(1-\hat{\eta}\right)^2} + \frac{\alpha}{1-\hat{\eta}}, \tag{72}$$

where R is a constant that can be pinned down using an initial condition. Suppose that at $t = 0$, the equilibrium is such that $\hat{\eta} = \hat{\eta}_0$. Then,

$$R = \frac{1-\alpha}{2\left(1-\hat{\eta}_0\right)^2} + \frac{\alpha}{\left(1-\hat{\eta}_0\right)} > 0.$$

Equation (72) provides the evolution of $\hat{\eta}$ over time. Since $\gamma > \sigma^2/2$ by assumption, it shows that $\hat{\eta} \to 1$ as $t \to \infty$. Combining (72) with 71 yields

$$\frac{dy}{dt} = \frac{g_\mu}{1-\alpha} \times \left(1 - \frac{1}{\sqrt{1 + \frac{1}{\gamma}\frac{1-\alpha}{\alpha}\left(\frac{g_\mu}{1-\varphi}t + -\frac{1-\alpha}{\alpha}a'\left(\hat{\eta}_0\right) - 2a\left(\hat{\eta}_0\right)\right)}}\right) > 0.$$

This expression can be written as (35). □

C.18 Proof of Corollary 2

Corollary 2. *For any $t > 0$, GDP grows faster in the economy with endogenous returns to scale. In the limit as $t \to \infty$, the long-run growth rates satisfy*

$$\lim_{t\to\infty} \frac{dy}{dt} = \frac{1}{1-\alpha}g_\mu > \frac{1}{1-\hat{\eta}_0\alpha}g_\mu = \lim_{t\to\infty}\frac{d\tilde{y}}{dt},$$

where $\tilde{y}$ is log GDP in the fixed returns-to-scale economy, and where $\hat{\eta}_0$ is effective returns to scale in the baseline economy at $t = 0$.

Proof. In the economy with exogenous returns to scale, (20) implies that $\frac{dy}{dt} = \frac{1}{1-\alpha\eta_0}g_\mu$. In the economy with endogenous returns to scale, the envelope theorem implies at, at any point in time we have $\frac{dy}{dt} = \frac{1}{1-\alpha\eta(t)}g_\mu$. This implies that the two economies have the same growth rate at $t = 0$ since $\eta(t) = \eta_0$ by definition. But since $d\eta/dt > 0$ by Corollary 1, the growth rate of the economy with endogenous returns to scale is larger for any $t > 0$. The second part of the result follows from taking the limit $t \to \infty$ in (35). □

D Robustness, extensions, and additional analysis

In this online supplement, we provide additional analysis of the benchmark model presented in the main text. We also show that that model can be extended in different ways.

D.1 Truncated normal shocks

In the baseline model, we assume that productivity shocks ε_{il} follow a normal distribution. While this allows for a tractable analytical solution, it theoretically permits firms to draw arbitrarily low productivity shocks, which could imply returns to scale $\eta_{il} \notin (0,1)$. In this online supplement, we solve the model assuming that productivity follows a *Truncated Normal* distribution. We show that the equilibrium conditions converge to those of the baseline model as the truncation point goes to negative infinity.

Specifically, productivity shocks of firms in industry i follow truncated normal distribution with support $[\underline{\varepsilon_i}, \infty)$. We assume that $\underline{\varepsilon_i}$ is sufficiently high such that

$$\underline{\varepsilon_i} > 2\gamma_i - s_i,$$

where s_i is given in (62). Under this restriction, all firms choose $\eta_{il} \in (0,1)$, as is evident from (62). The analogue of the free-entry condition, given by (8) in the main text, is

$$\exp\left(-s_i\right) \int_{\underline{\varepsilon_i}}^{\infty} \exp\left\{\frac{\varepsilon_{il}^2 - s_i^2}{4\gamma_i} + \frac{s_i^2 + \varepsilon_{il} s_i}{2\gamma_i}\right\} \exp\left(-\frac{\left(\varepsilon_{il} - \mu_i\right)^2}{2\sigma_i^2}\right) \frac{1}{1 - \Phi\left(\frac{\underline{\varepsilon_i} - \mu_i}{\sigma_i}\right)} d\varepsilon_{il} = \kappa_i \frac{W}{\lambda_i}.$$

Simplifying this expression, we get

$$\frac{W}{\lambda_i}\kappa_i = \frac{W^{1-\sum_j \alpha_{ij}} \prod_{j=1}^{N} P_j^{\alpha_{ij}}}{\lambda_i} \sqrt{\frac{1}{1-\varphi_i}} \exp\left(\frac{1}{2} \frac{\left(s_i + \mu_i\right)^2}{2\gamma_i \left(1 - \varphi_i\right)}\right) T_{1i}, \tag{73}$$

where, as above, $\varphi_i = \frac{\sigma_i^2}{2\gamma_i}$, and

$$T_{1i} = \frac{1 - \Phi\left(\frac{\sqrt{1-\varphi_i}}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i - \frac{\varphi_i}{1-\varphi_i}\left(\mu_i + s_i\right)\right)\right)}{1 - \Phi\left(\frac{1}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i\right)\right)}.$$

In the baseline model, $\underline{\varepsilon_i} = -\infty$, and $T_{1i} = 1$.

Next, following the same steps as in Supplement C.8, we can compute $\hat{\eta}_i$:

$$\hat{\eta}_i = \frac{\int_l \eta_{il} P_i Q_{il} dl}{\int_l P_i Q_{il} dl} = 1 - \frac{1-\varphi_i}{\frac{s_i+\mu_i}{2\gamma_i} + (1-\varphi_i) T_2} = 1 - \frac{(1-\varphi_i)(1-\eta_i(\mu_i))}{1+(1-\varphi_i)(1-\eta_i(\mu_i)) T_{2i}},$$

where

$$T_{2i} = \frac{\frac{1}{2\gamma_i}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\sqrt{1-\varphi_i}}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i - \frac{\varphi_i}{1-\varphi_i}(\mu_i + s_i)\right)\right)^2\right)}{1-\Phi\left(\frac{\sqrt{1-\varphi_i}}{\sigma_i}\left(\underline{\varepsilon_i} - \mu_i - \frac{\varphi_i}{1-\varphi_i}(\mu_i+s_i)\right)\right)}.$$

Clearly, if $\underline{\varepsilon_i} = -\infty$, then $T_{2i} = 0$, and we are back to the baseline model (see Equation (64)).

Finally, we can derive the analogue of (19). Following the same steps as in Supplement C.8, we get

$$\log\frac{P}{W} = -\left(I - [\operatorname{diag}(\varphi) + (I - \operatorname{diag}(\varphi))\operatorname{diag}(\eta(\mu))]\,\alpha\right)^{-1} \times \left[\mu + a(\eta(\mu)) - (I - \operatorname{diag}(\varphi))(I - \operatorname{diag}(\eta(\mu)))\left(\frac{1}{2}\log(1-\varphi) + \log\kappa - \log T_1\right)\right].$$

Again, if $\underline{\varepsilon_i} = -\infty$, we are back to the baseline model.

Clearly, if firms with $\varepsilon_{il} = \mu_i$ choose $\eta_{il} \in (0,1)$, $\underline{\varepsilon_i}$ can be chosen such that $\underline{\varepsilon_i} < \mu_i$. Furthermore, if σ_i is sufficiently small, T_{1i} is arbitrarily close to one and T_{2i} is arbitrarily close to zero. In that case, the mass of firms choosing $\eta_{il} \notin (0,1)$ in our baseline model is negligible, and the baseline economy is almost equivalent to the model with truncated normal shocks.

D.2 The impact of γ and κ on returns to scale

In this online supplement, we characterize how entry costs κ and the cost of scalability γ affect returns to scale.

D.3 Entry cost

We examine the impact of entry costs on returns-to-scale decisions.

Lemma 9. *The impact of the entry cost κ_j on the effective returns to scale $\hat{\eta}_j$ is given by*

$$\frac{d\hat{\eta}_i}{d\log\kappa_j} = \Psi_i^{-1}\left[-\mathcal{K}_{ij}(1-\hat{\eta}_j) - \mathbb{I}_{\{i=j\}}\right]. \tag{74}$$

In particular, $d\hat{\eta}_i/d\log\kappa_j \leq 0$ for $i \neq j$.

Proof. Applying the implicit function theorem to (68), we get

$$\frac{d\hat{\eta}}{d\log\kappa_j} = -\left[\frac{\partial\mathcal{F}}{\partial\hat{\eta}}\right]^{-1}\frac{\partial\mathcal{F}}{\partial\log\kappa_j}.$$

We have already computed the first term in the proof of Lemma 6, so consider the second one. We have

$$\frac{\partial\mathcal{F}_i}{\partial\log\kappa_j} = \alpha_i^\top\mathcal{L}(\hat{\eta})\frac{\partial z(\hat{\eta})}{\partial\log\kappa_j} + \frac{\partial\log\kappa_i}{\partial\log\kappa_j} = -\alpha_i^\top\mathcal{L}(\hat{\eta})1_j(1-\hat{\eta}_j) + \mathbb{1}(i=j).$$

Putting the pieces together we find the result. □

An increase in the entry cost in sector j always reduces the effective returns to scale of any other sector $i \neq j$. The mechanism is similar to that of a shock to μ_j. Increasing κ_j decreases j's productivity z_j, which increases the price of the input bundle of any sector that relies on j. Firms in those sectors then reduce their returns to scale to rely less on expensive intermediate inputs. At the same time, the effective returns to scale $\hat{\eta}_j$ of sector j itself typically increases with κ_j. This is because, when entry costs are large, there is more pressure to have fewer but larger firms, which requires large $\hat{\eta}_j$.

D.4 Cost of adjusting returns to scale

The productivity cost γ_i of adjusting returns to scale also affects firms' scalability decisions.

Lemma 10. *The impact of the productivity cost of higher returns to scale γ_j on the effective returns to scale $\hat{\eta}_i$ is given by*

$$\frac{d\hat{\eta}_i}{d\gamma_j} = \Psi_i^{-1}\left(\mathcal{K}_{ij}\frac{\partial z_j}{\partial\gamma_j} - \mathbb{I}_{\{i=j\}}\frac{\partial^2 z_i}{\partial\gamma_i\partial\hat{\eta}_i}\right) \tag{75}$$

where $\frac{\partial z_j}{\partial\gamma_j} = -\frac{1}{1-\hat{\eta}_j} - \frac{1}{2\gamma_j}\frac{\varphi_j}{1-\varphi_j}(1-\hat{\eta}_j) < 0$ *and* $\frac{\partial^2 z_i}{\partial\gamma_i\partial\hat{\eta}_i} = -\frac{1}{(1-\hat{\eta}_i)^2} + \frac{1}{2\gamma_i}\frac{\varphi_i}{1-\varphi_i}$. *In particular, $d\hat{\eta}_i/d\gamma_j \leq 0$ for $i \neq j$.*

Proof. Applying the implicit function theorem to (68), we get

$$\frac{d\hat{\eta}}{d\gamma_j} = -\left[\frac{\partial\mathcal{F}}{\partial\hat{\eta}}\right]^{-1}\left[\frac{\partial\mathcal{F}}{\partial\gamma_j}\right].$$

We have already computed the first term in the proof of Lemma 6, so consider the second one. We have

$$\frac{\partial\mathcal{F}_i}{\partial\gamma_j} := \alpha_i^\top\mathcal{L}(\hat{\eta})\frac{\partial z(\hat{\eta})}{\partial\gamma_j} + \frac{\partial^2 a_i(\hat{\eta}_i)}{\partial\gamma_j d\hat{\eta}_i} + \frac{1}{2}\frac{\partial}{\partial\gamma_j}\log(1-\varphi_i).$$

If $i \neq j$,

$$\frac{\partial \mathcal{F}_i}{\partial \gamma_j} := -\mathcal{K}_{ij} \frac{\partial z_j}{\partial \gamma_j},$$

where

$$\frac{\partial z_j}{\partial \gamma_j} = -\frac{1}{1-\hat{\eta}_j} - \frac{1}{2\gamma_j} \frac{\varphi_j}{1-\varphi_j} \left(1-\hat{\eta}_j\right) < 0.$$

For $i = j$, we have an extra term,

$$\frac{\partial \mathcal{F}_i}{\partial \gamma_i} := -\mathcal{K}_{ii} \frac{\partial z_i\left(\hat{\eta}\right)}{\partial \gamma_i} - \frac{1}{\left(1-\hat{\eta}_i\right)^2} + \frac{1}{2\gamma_i} \frac{\varphi_i}{1-\varphi_i}.$$

□

Consider first the impact of a higher γ_j on the effective returns to scale of another sector $i \neq j$. Unsurprisingly, a higher productivity cost of adjusting returns to scale leads to a lower productivity in sector j. Through input-output linkages, that lower productivity increases the price of the intermediate input bundles of firms that rely, directly or indirectly, on j as an input ($\mathcal{L}_{ij} > 0$). Those firms, to limit the negative impact of higher inputs, lower their returns to scale. A similar impact is at work when considering the impact of a higher γ_j on j itself, but in addition, j is also affected more directly by the increase in γ_j. Indeed, a larger γ_j mechanically makes a high $\hat{\eta}_j$ more expensive, which amplifies the negative movement in $\hat{\eta}_j$. In general, these forces combine to create a stronger negative impact of γ_j on $\hat{\eta}_j$.

D.5 Wedges

In this online supplement, we consider an economy with wedges. In the presence of wedges, the firm's problem (2) becomes

$$\Pi_{il} := \max_{\eta_{il}, L_{il}, X_{il}} \left(1-\tau_i^S\right) P_i F_i\left(L_{il}, X_{il}, \eta_{il}\right) - \left(1+\tau_i^L\right) W L_{il} - \sum_{j=1}^{N} \left(1+\tau_{ij}^X\right) P_j X_{ij,l}. \tag{76}$$

Firms in sector i have to pay $\left(1+\tau_{ij}^X\right) P_j$ for each unit of good j, $\tau_{ij}^X > -1$, and $\left(1+\tau_i^L\right) W$ for each unit of labor, $\tau_i^L > -1$. Firms face an effective sales tax $\tau_i^S < 1$. Finally, we introduce a corporate tax rate τ_i^Π. This tax does not directly affect the profit-maximization problem (76). However, it affects the free-entry condition (8):

$$\mathrm{E}_i\left[\left(1-\tau_i^\Pi\right) \Pi_i\left(\varepsilon_{il}, P^*, W^*\right)\right] = \kappa_i W^*.$$

As we can see, the profit tax effectively increases the entry cost.

Wedges $\left\{\tau^X, \tau^L, \tau^S, \tau^\Pi\right\}$ can capture a variety of economic factors, such as tariffs, transportation costs, taxes, markups, etc. Some of those wedges can be associated with loss of resources, while others only lead to resource redistribution. To capture this, we assume that a fraction of wedge income is rebated to the household, such that its budget constraint (7) becomes

$$\sum_{i=1}^{N} P_i C_i \leq W\bar{L} + \mathcal{T},$$

where

$$\mathcal{T} = \sum_{i=1}^{N} \theta_i^S \tau_i^S P_i Q_i + \sum_{i=1}^{N} \theta_i^L \tau_i^L W L_i + \sum_{i=1}^{N}\sum_{j=1}^{N} \theta_{ij}^X \tau_{ij}^X P_j X_{ij} + \sum_{i=1}^{N} \theta_i^\Pi \tau_i^\Pi \Pi_i.$$

Here $\theta_i^S, \theta_i^L, \theta_{ij}^X, \theta_i^\Pi \in [0,1]$. Note that wedges $\left\{\tau^X, \tau^L, \tau^S, \tau^\Pi\right\}$ can be both positive or negative. For example, τ_{ij}^X is positive in case of transportation costs. If those are iceberg costs, nothing is rebated to the household, and $\theta_{ij}^X = 0$. Tariffs would also correspond to a positive τ_{ij}^X. Different from transportation costs, tariff income is likely partially rebated to the household, in which case θ_{ij}^X is positive. On the other hand, τ_{ij}^X would be negative in case of government subsidies. Such subsidies are financed by lump-sum taxation of the household, such that $\theta_{ij}^X = 1$.[58]

The model can be analyzed analogously to our baseline model. In particular, we can derive that the equilibrium price vector is given by

$$\log \frac{P}{W} = -\mathcal{L}\left(\hat{\eta}\right) z\left(\hat{\eta}\right), \tag{77}$$

where, as in the baseline model, $\hat{\eta}$ is a vector of sales-weighted average returns to scale, $\mathcal{L}\left(\hat{\eta}\right) = \left(I - \operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}$, and

$$z_i\left(\hat{\eta}_i\right) = \mu_i - T_i + a_i\left(\hat{\eta}_i\right) + \frac{\sigma_i^2}{2}\frac{1}{1-\hat{\eta}_i} + \frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\varphi_i}\right) - \left(1-\hat{\eta}_i\right)\log\frac{\kappa_i}{1-\tau_i^\Pi}. \tag{78}$$

The productivity shifter T_i is

$$T_i = T_i\left(\tau_i^L, \tau_i^S, \tau_i^X, \hat{\eta}_i\right) = \log\left[\frac{\left(1+\tau_i^L\right)^{\hat{\eta}_i\left(1-\sum_j \alpha_{ij}\right)}\prod_{j=1}^{N}\left(1+\tau_{ij}^X\right)^{\hat{\eta}_i\alpha_{ij}}}{1-\tau_i^S}\right]. \tag{79}$$

[58]Deadweight losses of subsidies can be captured by setting $\theta_{ij}^X > 1$.

Introducing wedges $\left\{\tau^X, \tau^L, \tau^S\right\}$ is, therefore, equivalent to a change in sectoral total factor productivities. An increase in wedges τ_i^L, τ_i^S or τ_{ij}^X reduces the effective productivity of sector i, resulting in a reduction in the returns to scale in all sectors. This result is analogous to the effect of a reduction in μ_i, described in Corollary 6. At the same time, an increase in the corporate tax τ_i^Π effectively increases the entry cost, and so its impact on returns to scale is analogous to that of $\log \kappa_i$, described in Corollary 9.

Proposition 10. *An increase in wedges $\left\{\tau^X, \tau^L, \tau^S\right\}$ reduces returns to scale in all sectors. An increase in the profit tax τ_i^Π reduces returns to scale in other sectors but can increase returns to scale in sector i.*[59]

The market-clearing conditions (9) also change. Specifically, for good i, the resource constraint becomes

$$\left(1 - \left(1 - \theta_i^S\right) \tau_i^S\right) Q_i = C_i + \sum_{j=1}^{N} \left(1 + \left(1 - \theta_{ji}^X\right) \tau_{ji}^X\right) X_{ji}.$$

Then the Domar weight of sector i is

$$\tilde{\omega}_i = \frac{P_i Q_i}{\bar{P} Y} = \mathbf{1}_i \left(I - \text{diag}\left[\left(1 - \theta^S\right) \circ \tau^S\right] - \tilde{\alpha}^\top \text{diag}\left(\hat{\eta}\right)\right)^{-1} \beta,$$

where $\circ$ denotes element-wise product of two vectors, and

$$\tilde{\alpha}_{ji} = \alpha_{ji} \left(1 - \tau_j^S\right) \left(\frac{1 + \left(1 - \theta_{ji}^X\right) \tau_{ji}^X}{1 + \tau_{ji}^X}\right).$$

Clearly, if $\tau_j^S \geq 0$ and $\tau_{ji}^X \geq 0$, then $\tilde{\alpha}_{ji} \leq \alpha_{ji}$.

Using these results, we can derive how wedges affect the expression for the aggregate output.

Proposition 11. *Equilibrium log GDP $y := \log Y$ is given by*

$$y\left(\hat{\eta}\right) = \underbrace{\beta^\top \mathcal{L}\left(\hat{\eta}\right) z\left(\hat{\eta}\right)}_{\textit{Contribution of productivity}} + \underbrace{\log \bar{L}}_{\textit{Labor endowment}} - \underbrace{\log \Gamma_\tau}_{\textit{Wedges income}}, \tag{80}$$

[59] We provide expressions for derivatives of returns to scale with respect to wedges in the proof of this proposition.

where

$$\Gamma_\tau =$$

$$1-\sum_{i=1}^{N}\tilde{\omega}_i\left(\left(1-\tau_i^S\right)\hat{\eta}_i\left(\sum_{j=1}^{N}\alpha_{ij}\frac{\theta_{ij}^X\tau_{ij}^X}{1+\tau_{ij}^X}+\left(1-\sum_{j=1}^{N}\alpha_{ij}\right)\frac{\theta_i^L\tau_i^L}{1+\tau_i^L}\right)+\theta_i^S\tau_i^S+\theta_i^\Pi\tau_i^\Pi\left(1-\tau_i^S\right)\left(1-\hat{\eta}_i\right)\right).$$

As discussed above, some wedges can lead to a destruction of resources while others may lead to redistribution of resources. In the latter case, aggregate output needs to be adjusted for wedges income. This is the last term in expression (80). Naturally, if $\theta_{ij}^X = \theta_i^L = \theta_i^S = \theta_i^\Pi = 0$ for all i and j, then nothing is rebated to the household, and $\log\Gamma_\tau = 0$. If all the wedges are nonnegative, and some of the wedge income is rebated back to the household, then $\log\Gamma_\tau < 0$, which leads to a higher y.[60]

The presence of wedges distorts the economy. Intuitively, firms do not internalize that part of the wedge income is rebated to the household, and their decisions are inefficient as a result. If none of the wedge income is rebated to the household, then $\log\Gamma_\tau = 0$, and the economy is efficient. In that case, firms correctly perceive wedges as resource-destructive. In the inefficient economy, the equilibrium returns to scale do not maximize GDP, and any marginal change in returns to scale can have a nontrivial impact on GDP. Specifically, a change in the underlying parameter χ leads to the following response of GDP:

$$\frac{dy}{d\chi}=\frac{\partial y}{\partial\chi}+\sum_{j=1}^{N}\frac{\partial y}{\partial\hat{\eta}_j}\frac{d\hat{\eta}_j}{d\chi}.$$

In general, the sign of the response of GDP to a marginal change in returns to scale, $\frac{\partial y}{\partial\hat{\eta}_j}$, depends on the sign of wedges. However, we can provide a sharp characterization in a few important special cases.

Proposition 12. *Suppose that there are no profit taxes,* $\tau_i^\Pi = 0$*, and all other wedges are positive,* $\tau_{ij}^X > 0$, $\tau_i^L > 0$*, and* $\tau_i^S > 0$ *for all* i, j*, and suppose that some of the wedge income is rebated to the household,* $\log\Gamma_\tau < 0$*. Then any marginal increase in the returns to scale leads to an increase in GDP,* $\frac{\partial y}{\partial\hat{\eta}_j} > 0$.

Consider first the case with no profit taxes. If other wedges are positive, the equilibrium returns to scale are too low (Proposition 10) as the firms do not internalize that part of the wedge income is

[60] Of course, in that case, sectoral productivities (78) are also lower than in the no-wedges economy.

rebated to the household. Then, any change in the parameter that leads to an increase in returns to scale is beneficial for GDP. For example, if the economy becomes more productive, as captured by a higher μ_j, the equilibrium returns to scale increase (Corollary 6).[61] Such a change has a positive impact on GDP because the equilibrium returns to scale were inefficiently low before the change.

Profit taxes affect the equilibrium returns to scale differently. As Proposition 10 suggests, an increase in the profit tax τ_i^Π is equivalent, from the firms' perspective, to an increase in the entry cost κ_i. Such an increase typically leads to a higher $\hat{\eta}_i$ (see our discussion following Corollary 9). Therefore, if profit taxes are rebated to the household, equilibrium returns to scale tend to be inefficiently high as firms incorrectly perceive entry costs as being too high. In that case, any change in the parameter that leads to a further increase in returns to scale is harmful for GDP. Expression (85) in the proof of Proposition 12 provides an exact expression for $\frac{\partial y}{\partial \hat{\eta}_j}$ in that case.

D.5.1 Proof of Proposition 10

Proposition 10. *An increase in wedges $\left\{\tau^X, \tau^L, \tau^S\right\}$ reduces returns to scale in all sectors. An increase in the profit tax τ_i^Π reduces returns to scale in other sectors but can increase returns to scale in sector i.*

Proof. Taking first-order conditions of (76) with respect to L_{il} and X_{il}, we can derive the following expression for log profit of firm l in sector i:

$$\log \Pi_{il} = \log P_i + \log\left(1-\tau_i^S\right) + \frac{a_i\left(\eta_{il}\right) + \eta_{il}\left(\log \frac{P_i}{W} - \sum_{j=1}^N \alpha_{ij} \log \frac{P_j}{W}\right)}{1-\eta_{il}}$$
$$+ \frac{\varepsilon_{il} - \eta_{il}\left(\log \frac{\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}}{\left(1-\tau_i^S\right)}\right)}{1-\eta_{il}}.$$

Then the first-order condition with respect to η_{il} yields

$$\frac{d \log \Pi_{il}}{d\eta_{il}} = 0 \Leftrightarrow \tag{81}$$
$$\varepsilon_{il} - \log\left[\frac{\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}}{\left(1-\tau_i^S\right)}\right] + \log \frac{P_i}{W} - \sum_{j=1}^N \alpha_{ij} \log \frac{P_j}{W} + a_i\left(\eta_{il}\right) + \left(1-\eta_{il}\right)\frac{da_i}{d\eta_{il}} = 0.$$

[61] If $\chi \in \left\{\mu_j, \sigma_j^2, \kappa_j, \gamma_j\right\}$, it is straightforward to show that $\frac{\partial y}{\partial \chi}$ and $\frac{d\hat{\eta}_j}{d\chi}$ are given by the same expressions as in the baseline model.

Following the same steps as in the proof of Proposition 2, we can derive that the equilibrium price vector is given by (77), and the sales-weighted average of firm-level returns to scale $\hat{\eta}_i$ satisfies

$$\frac{1}{1-\eta_{il}} = \frac{1-\varphi_i}{1-\hat{\eta}_i} + \frac{\varepsilon_{il} - \tilde{\mu}_i}{2\gamma_i}, \tag{82}$$

where $\tilde{\mu}_i = \mu_i - \log\left[\frac{\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}}{\left(1-\tau_i^S\right)}\right]$. Plugging (77) and (82) into (81), we get the following equation for $\hat{\eta}$:

$$\begin{aligned}\mathcal{F}_i =& \frac{da_i\left(\hat{\eta}_i\right)}{d\hat{\eta}_i} + \frac{\sigma_i^2}{2}\frac{1}{\left(1-\hat{\eta}_i\right)^2} + \alpha_i^\top \mathcal{L}\left(\hat{\eta}\right) z\left(\hat{\eta}\right) + \frac{1}{2}\log\left(1-\varphi_i\right) + \log\frac{\kappa_i}{1-\tau_i^\Pi} \\ &- \log\left[\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}\right] = 0.\end{aligned} \tag{83}$$

Denote by χ_i any of τ_{ij}^X, τ_i^L or τ_i^S. Then, by the implicit function theorem,

$$\frac{d\hat{\eta}}{d\chi_i} = -\left[\frac{\partial\mathcal{F}}{\partial\hat{\eta}}\right]^{-1}\left[\frac{\partial\mathcal{F}}{\partial\chi_i}\right].$$

As in the baseline model, we have

$$\frac{\partial\mathcal{F}_i}{\partial\hat{\eta}_i} = \left(1-\varphi_i\right)\frac{d^2a_i}{d\hat{\eta}_i^2}$$

and $\frac{\partial\mathcal{F}_i}{\partial\hat{\eta}_j} = 0$ if $i \neq j$. Furthermore,

$$\frac{\partial\mathcal{F}_i}{\partial\chi_k} = \alpha_i^\top \mathcal{L}\left(\hat{\eta}\right)\frac{\partial z\left(\hat{\eta}\right)}{\partial\chi_k} - \frac{\partial\log\left[\left(1+\tau_i^L\right)^{1-\sum_{j=1}^N \alpha_{ij}} \prod_{j=1}^N \left(1+\tau_{ij}^X\right)^{\alpha_{ij}}\right]}{\partial\chi_k}.$$

From (78), it is clear that $\frac{\partial z_k}{\partial\chi_k} < 0$. Therefore, $\frac{d\hat{\eta}}{d\chi_i} \leq 0$. In particular, we have

$$\frac{d\hat{\eta}_i}{d\tau_j^S} = -\frac{1}{1-\tau_j^S}\frac{d\hat{\eta}_i}{d\log\left(1-\tau_j^S\right)} = -\frac{1}{1-\tau_j^S}\left[\left(1-\varphi_i\right)\frac{d^2a_i}{d\hat{\eta}_i^2}\right]^{-1}\mathcal{K}_{ij}.$$

□

D.5.2 Proof of Proposition 11

Proposition 11. *Equilibrium log GDP* $y := \log Y$ *is given by*

$$y(\hat{\eta}) = \underbrace{\beta^{\top}\mathcal{L}(\hat{\eta})\,z(\hat{\eta})}_{\text{Contribution of productivity}} + \underbrace{\log \bar{L}}_{\text{Labor endowment}} - \underbrace{\log \Gamma_{\tau}}_{\text{Wedges income}},$$

where

$$\Gamma_{\tau} =$$
$$1-\sum_{i=1}^{N}\tilde{\omega}_i\left(\left(1-\tau_i^S\right)\hat{\eta}_i\left(\sum_{j=1}^{N}\alpha_{ij}\frac{\theta_{ij}^X\tau_{ij}^X}{1+\tau_{ij}^X}+\left(1-\sum_{j=1}^{N}\alpha_{ij}\right)\frac{\theta_i^L\tau_i^L}{1+\tau_i^L}\right)+\theta_i^S\tau_i^S+\theta_i^{\Pi}\tau_i^{\Pi}\left(1-\tau_i^S\right)\left(1-\hat{\eta}_i\right)\right).$$

Proof. From the household's budget constraint, we have

$$Y = W\bar{L}+\mathcal{T}, \tag{84}$$

where

$$\begin{aligned}
\mathcal{T} &= \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\theta_{ij}^X\tau_{ij}^X P_j X_{ij}+\theta_i^L\tau_i^L W L_i+\theta_i^S\tau_i^S P_i Q_i+\theta_i^{\Pi}\tau_i^{\Pi}\Pi_i\right)\\
&= \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\hat{\eta}_i\alpha_{ij}\frac{\theta_{ij}^X\tau_{ij}^X}{1+\tau_{ij}^X}\left(1-\tau_i^S\right)P_iQ_i+\hat{\eta}_i\left(1-\sum_{j=1}^{N}\alpha_{ij}\right)\frac{\theta_i^L\tau_i^L}{1+\tau_i^L}\left(1-\tau_i^S\right)P_iQ_i\right.\\
&\left.+\theta_i^S\tau_i^S P_iQ_i+\theta_i^{\Pi}\tau_i^{\Pi}\left(1-\tau_i^S\right)\left(1-\hat{\eta}_i\right)P_iQ_i\right)\\
&= Y\sum_{i=1}^{N}\tilde{\omega}_i\left(\left(1-\tau_i^S\right)\hat{\eta}_i\left(\sum_{j=1}^{N}\frac{\alpha_{ij}\theta_{ij}^X\tau_{ij}^X}{1+\tau_{ij}^X}+\frac{\left(1-\sum_{j=1}^{N}\alpha_{ij}\right)\theta_i^L\tau_i^L}{1+\tau_i^L}\right)+\theta_i^S\tau_i^S+\theta_i^{\Pi}\tau_i^{\Pi}\left(1-\tau_i^S\right)\left(1-\hat{\eta}_i\right)\right).
\end{aligned}$$

Plugging this into (84) gives the result. □

D.5.3 Proof of Proposition

Proposition 10. *Suppose that there are no corporate taxes,* $\tau_i^{\Pi}=0$*, and all other wedges are positive,* $\tau_{ij}^X>0$*,* $\tau_i^L>0$*, and* $\tau_i^S>0$ *for all* i,j*, and suppose that some of the wedge income is rebated to the household,* $\log\Gamma_{\tau}<0$*. Then any marginal increase in the returns to scale leads to an* *increase in GDP,* $\frac{\partial y}{\partial \hat{\eta}_j} > 0$.

Proof. Differentiating y, given by (80), with respect to $\hat{\eta}_i$ and using the first-order condition (83), we get

$$\frac{\partial y}{\partial \hat{\eta}_i} = -\frac{\partial \log \Gamma_\tau}{\partial \hat{\eta}_i} = \frac{\text{Num}_i}{\Gamma_\tau},$$

where the numerator is

$$\text{Num}_i = \tilde{\omega}_i \left(1 - \tau_i^S\right) \left(\sum_{j=1}^N \alpha_{ij} \frac{\theta_{ij}^X \tau_{ij}^X}{1 + \tau_{ij}^X} + \left(1 - \sum_{j=1}^N \alpha_{ij}\right) \frac{\theta_i^L \tau_i^L}{1 + \tau_i^L} - \theta_i^\Pi \tau_i^\Pi \right)$$
$$+ \sum_{k=1}^N \frac{d\tilde{\omega}_k}{d\hat{\eta}_i} \left(\left(1 - \tau_k^S\right) \hat{\eta}_k \left(\sum_{j=1}^N \alpha_{kj} \frac{\theta_{kj}^X \tau_{kj}^X}{1 + \tau_{kj}^X} + \left(1 - \sum_{j=1}^N \alpha_{kj}\right) \frac{\theta_k^L \tau_k^L}{1 + \tau_k^L} \right) + \theta_k^S \tau_k^S + \theta_k^\Pi \tau_k^\Pi \left(1 - \tau_k^S\right) \left(1 - \hat{\eta}_k\right) \right).$$

The derivative of the Domar weights is given by

$$\frac{d\tilde{\omega}_k}{d\hat{\eta}_i} = \sum_{j=1}^N \tilde{\alpha}_{ij} \left(I - \text{diag}\left[\left(1 - \theta^S\right) \circ \tau^S\right] - \text{diag}\left(\hat{\eta}\right) \tilde{\alpha}\right)_{jk}^{-1} \tilde{\omega}_i > 0.$$

Therefore, if taxes are positive ($\tau^X, \tau^L, \tau^S > 0$) but there is no profit tax ($\tau^\Pi = 0$), all terms in the numerator are positive (assuming some rebates $\theta > 0$), implying $\frac{\partial y}{\partial \hat{\eta}_i} > 0$. □

In contrast, if $\tau^X = \tau^L = \tau^S = 0$ and $\tau^\Pi > 0$ with $\theta^\Pi = 1$, then $\tilde{\omega} = \omega$, and we get

$$\frac{\partial y}{\partial \hat{\eta}_i} = \omega_i \frac{-\tau_i^\Pi + \sum_{k=1}^N \left(\sum_{j=1}^N \alpha_{ij} \mathcal{L}_{jk}\right) \tau_k^\Pi \left(1 - \hat{\eta}_k\right)}{1 - \sum_{i=1}^N \omega_i \tau_i^\Pi \left(1 - \hat{\eta}_i\right)}. \tag{85}$$

As $\hat{\eta}_k \to 1$ for all k, the term $(1 - \hat{\eta}_k)$ vanishes, leaving only the negative term $-\omega_i \tau_i^\Pi$. Thus, $\frac{\partial y}{\partial \hat{\eta}_i}$ becomes negative.

D.6 Sales wedge correlated with productivity

In this online supplement, we consider an economy in which firms face sales tax (37). The firm's problem (2) becomes

$$\Pi_{il} := \max_{\eta_{il}, L_{il}, X_{il}} \left(1 - \tau_{il}^S\right) P_i F_i \left(L_{il}, X_{il}, \eta_{il}\right) - W L_{il} - \sum_{j=1}^N P_j X_{ij,l},$$

where $F_i\left(L_{il}, X_{il}, \eta_{il}\right)$ is given by (1). Clearly, this problem is equivalent to the one in the main text if we redefine the productivity as

$$\tilde{\varepsilon}_{il} = \varepsilon_{il} + \log\left(1-\tau_{il}^S\right) = \left(1-b_i\right)\left(\varepsilon_{il}-\mu_i\right) + \mu_i + \log\left(1-\tau_i^S\right),$$

such that $\tilde{\varepsilon}_{il} \sim \text{iid}\ \mathcal{N}\left(\tilde{\mu}_i, \tilde{\sigma}_i^2\right)$, where

$$\tilde{\mu}_i = \mu_i + \log\left(1-\tau_i^S\right) \quad \text{and} \quad \tilde{\sigma}_i = \left(1-b_i\right)\sigma_i.$$

Similar to the baseline model, define

$$\hat{\eta}_i := \int_0^{M_i} \frac{\tilde{\omega}_{il}}{\tilde{\omega}_i}\eta_{il} dl, \tag{86}$$

with $\tilde{\omega}_{il} := \left(1-\tau_{il}\right)\omega_{il}$ and $\tilde{\omega}_i := \left(1-\hat{\tau}_i^S\right)\omega_i$, where

$$\hat{\tau}_i^S = \int_0^{M_i} \frac{\omega_{il}}{\omega_i}\tau_{il}^S dl. \tag{87}$$

As in the baseline model, we get

$$\frac{1}{1-\hat{\eta}_i} = \frac{\tilde{\mu}_i + s_i}{2\gamma_i\left(1-\tilde{\varphi}_i\right)},$$

where $s_i = \log P_i - \log H_i$. Integrating (87), we get

$$\frac{1}{1-\hat{\tau}_i^S} = \frac{1}{1-\tau_i^S}\left(1 + \frac{\tilde{\varphi}_i}{\frac{1-\tilde{\varphi}_i}{1-\hat{\eta}_i}}\frac{b_i}{1-b_i}\right)\exp\left(-\frac{b_i}{1-b_i}\frac{4\tilde{\varphi}_i\gamma_i\frac{1-\tilde{\varphi}_i}{1-\hat{\eta}_i} + \tilde{\sigma}_i^2\frac{b_i}{1-b_i}}{2}\frac{1}{\left(1-\tilde{\varphi}_i\right)}\right). \tag{88}$$

Then, following the same steps as in the main model, we can derive

$$\log W = \beta^\top\left(I - \operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1}\tilde{z}\left(\hat{\eta}\right),$$

where

$$\tilde{z}_i\left(\hat{\eta}_i\right) = \tilde{\mu}_i + a_i\left(\hat{\eta}_i\right) + \frac{\tilde{\sigma}_i^2}{2}\frac{1}{1-\hat{\eta}_i} + \frac{1}{2}\left(1-\hat{\eta}_i\right)\log\left(\frac{1}{1-\tilde{\varphi}_i}\right) - \left(1-\hat{\eta}_i\right)\log\kappa_i$$

and $\tilde{\varphi}_i = \frac{\tilde{\sigma}_i^2}{2\gamma_i}$.

We assume that all tax proceeds are rebated to the household. Therefore, using the market clearing condition (9), we get

$$\omega_i := \frac{P_i Q_i}{\bar{P} Y} = \beta^\top \left(I - \text{diag}\left(1 - \hat{\tau}^S\right) \text{diag}\left(\hat{\eta}\right) \alpha\right) \mathbf{1}_i.$$

The rebate amount is

$$\mathcal{T} = \sum_{i=1}^{N} \int_0^{M_i} \tau_{il}^S P_i Q_{il} dl = \sum_{i=1}^{N} \hat{\tau}_i^S \omega_i \bar{P} Y.$$

GDP is then

$$C = W\bar{L} + \mathcal{T} \Leftrightarrow \log C = \log W + \log \bar{L} - \log\left(1 - \sum_{i=1}^{N} \hat{\tau}_i^S \omega_i\right).$$

D.7 Dispersed returns-to-scale economy

In this online supplement, we consider the dispersed returns-to-scale economy. Specifically, consider the initial economy (we will use subscripts b to mark any quantities in that economy). From (10), firm l in sector i chooses the following returns to scale:

$$\frac{1}{1 - \eta_{il}^b} = \frac{1}{2\gamma_i}\left(\varepsilon_{il}^b + s_i^b\right), \tag{89}$$

where $s_i^b = \log P_i^b - \log H_i^b$ in the initial economy. Furthermore, from (17), we know that

$$\frac{1}{1 - \hat{\eta}_i^b} = \frac{1}{2\gamma_i\left(1 - \varphi_i^b\right)}\left(\mu_i^b + s_i^b\right), \tag{90}$$

where $\varphi_i^b = \frac{\left(\sigma_i^b\right)^2}{2\gamma_i}$.

Suppose now that there is a change in the distribution of ε_{il}^b, such that the mean changes from μ_i^b to μ_i, and the standard deviation changes from σ_i^b to σ_i. Such a change can reflect an increase in μ for all sectors (Section 7.2) or removal of sales tax (Section 7.3). Then, productivity ε_{il}^b shifts to ε_{il}, where

$$\frac{\varepsilon^b - \mu_i^b}{\sigma_i^b} = \frac{\varepsilon_{il} - \mu_i}{\sigma_i}.$$

In the dispersed economy, firms can adjust all their choices except returns to scale. The free-entry condition (8)

$$\underbrace{\int_{-\infty}^{\infty} \exp\left(\log P_i + \frac{\varepsilon_{il} + a_i\left(\eta_{il}^b\right) + \eta_{il}^b s_i}{1-\eta_{il}^b}\right)}_{=\Pi_{il}\left(\varepsilon_{il},\eta_{il}^b\right)} f_i\left(\varepsilon_{il}\right) d\varepsilon_{il} = \kappa_i W,$$

where $s_i = \log P_i - \log H_i$, and η_{il}^b is given by (89). Taking this integral, we get

$$\exp\left(\frac{\left[\left(\mu_i^b + s_i^b\right)\left(1-\varphi_i^b \frac{\sigma_i}{\sigma_i^b}\right) + \varphi_i^b\left(\mu_i + s_i\right)\right]^2 \frac{1}{1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)} - \left(\mu_i^b + s_i^b\right)^2}{2\left(\sigma_i^b\right)^2}\right)\times \tag{91}$$

$$\exp\left(\sum_{j=1}^{N} \alpha_{ij} \log \frac{P_j}{W}\right) \frac{1}{\sqrt{1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)}} = \kappa_i.$$

Next, we can define $\hat{\eta}_i$ in the same way as usual, $\hat{\eta}_i = \int_0^{M_i} \frac{\omega_{il}}{\omega_i}\eta_{il}^b dl$, where again η_{il}^b is given by (89). Omitting tedious yet straightforward calculations, we get

$$\frac{1}{1-\hat{\eta}_i} = \frac{1}{2\gamma_i\left(1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)\right)}\left[\left(\mu_i^b + s_i^b\right) + \varphi_i^b\left(\mu_i + s_i - \left(\mu_i^b + s_i^b\right)\frac{\sigma_i}{\sigma_i^b}\right)\right]. \tag{92}$$

Combining (90), (91), and (92), we get

$$\log W = \beta^\top \left(I - \operatorname{diag}\left(\hat{\eta}\right)\alpha\right)^{-1} z,$$

where

$$\begin{aligned} z_i =& \mu_i - \left[\frac{1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)}{1-\hat{\eta}_i} - 2\frac{1-\varphi_i^b}{1-\hat{\eta}_i^b}\left(1-\varphi_i^b\frac{\sigma_i}{\sigma_i^b}\right) + \left(1-\hat{\eta}_i\right)\left(\frac{1-\varphi_i^b}{1-\hat{\eta}_i^b}\right)^2\right]\frac{\gamma_i}{\varphi_i^b} \\ & - \left(1-\hat{\eta}_i\right)\left[\frac{1}{2}\log\left(1-\varphi_i^b\left(2\frac{\sigma_i}{\sigma_i^b}-1\right)\right) + \log\kappa\right]. \end{aligned}$$

GDP is then $Y = W\bar{L}$.

References

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer (2015). "Identification properties of recent production function estimators". *Econometrica* 83.6, pp. 2411–2451.

Aw, Bee Yan, Xiaomin Chen, and Mark J Roberts (2001). "Firm-level evidence on productivity differentials and turnover in Taiwanese manufacturing". *Journal of development economics* 66.1, pp. 51–86.

Baqaee, David Rezza and Emmanuel Farhi (2019). "Productivity and Misallocation in General Equilibrium". *Quarterly Journal of Economics* 135.1, pp. 105–163.

Blundell, Richard and Stephen Bond (2000). "GMM estimation with persistent panel data: an application to production functions". *Econometric reviews* 19.3, pp. 321–340.

Caves, Douglas W, Laurits R Christensen, and W Erwin Diewert (1982a). "Multilateral comparisons of output, input, and productivity using superlative index numbers". *Economic Journal* 92.365, pp. 73–86.

— (1982b). "The economic theory of index numbers and the measurement of input, output, and productivity". *Econometrica*, pp. 1393–1414.

Chiavari, Andrea (2024). "Customer Accumulation, Returns to Scale, and Secular Trends".

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger (2020). "The rise of market power and the macroeconomic implications". *Quarterly Journal of Economics* 135.2, pp. 561–644.

De Ridder, Maarten, Basile Grassi, and Giovanni Morzenti (2022). "The Hitchhiker's Guide to Markup Estimation".

Demirer, Mert (2025). *Production function estimation with factor-augmenting technology: An application to markups.* Working Paper.

Feenstra, Robert C, Robert Inklaar, and Marcel P Timmer (2015). "The next generation of the Penn World Table". *American economic review* 105.10, pp. 3150–3182.

Gabaix, Xavier and Rustam Ibragimov (2011). "Rank- 1/2: a simple way to improve the OLS estimation of tail exponents". *Journal of Business & Economic Statistics* 29.1, pp. 24–39.

Gale, David and Hukukane Nikaido (1965). "The Jacobian matrix and global univalence of mappings". *Mathematische Annalen* 159.2, pp. 81–93.

Gao, Wei and Matthias Kehrig (2025). *Returns to scale, productivity and competition: Empirical evidence from US manufacturing and construction establishments.* Working Paper.

Hsieh, Chang-Tai and Peter J Klenow (2009). "Misallocation and manufacturing TFP in China and India". *Quarterly Journal of Economics* 124.4, pp. 1403–1448.

Hubmer, Joachim, Mons Chan, Serdar Ozkan, Sergio Salgado, and Guangbin Hong (2025). *Scalable versus Productive Technologies.* Tech. rep.

Kalemli-Özcan, Şebnem, Bent E Sørensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yeşiltaş (2024). "How to Construct Nationally Representative Firm-Level Data from the Orbis Global Database: New Facts on SMEs and Aggregate Implications for Industry Concentration". *American Economic Journal: Macroeconomics* 16.2, pp. 353–374.

Kopytov, Alexandr, Mathieu Taschereau-Dumouchel, and Zebang Xu (2024). "The Origin of Risk". *Available at SSRN.*

Levinsohn, James and Amil Petrin (2003). "Estimating production functions using inputs to control for unobservables". *Review of Economic Studies* 70.2, pp. 317–341.

McAdam, Peter, Philipp Meinen, Chris Papageorgiou, and Patrick Schulte (2024). *Returns to scale: New evidence from administrative firm-level data.* Tech. rep. 24/2024.

Olley, G Steven and Ariel Pakes (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". *Econometrica* 64.6, pp. 1263–1297.

Ruzic, Dimitrije and Sui-Jade Ho (2023). "Returns to scale, productivity, measurement, and trends in US Manufacturing Misallocation". *Review of Economics and Statistics* 105.5, pp. 1287–1303.

Savagar, Anthony and Joel Kariel (2024). "Scale Economies and Aggregate Productivity".

Söderbom, Måns (2009). "Imposing common factor restrictions for AR (1) model" ex post" using a minimum distance procedure". *stata command, Downloaded from: http://www. soderbom. net/comfac_ md. zip.*

Star, Spencer and Robert E Hall (1976). "An approximate Divisia index of total factor productivity". *Econometrica: Journal of the Econometric Society*, pp. 257–263.

Teti, Feodora A. (2024). "Missing Tariffs".