Endogenous Returns to Scale

Alexandr Kopytov
University of Rochester

Mathieu Taschereau-Dumouchel
Cornell University

Zebang Xu
Cornell University

1910: Automobiles were luxury goods carefully assembled by skilled craftsmen (Packard Co.)


1910: Automobiles were luxury goods carefully assembled by skilled craftsmen (Packard Co.)
Good at small scale, hard to scale up

1913: Henry Ford's Highland Park plant adopts the assembly line and produces a Model T every 93 minutes


1913: Henry Ford’s Highland Park plant adopts the assembly line and produces a Model T every 93 minutes
Good at large scale, impractical at small scale

What determines returns to scale?

By adopting the moving assembly line, Ford *chose* higher return to scale

- Better ability to scale up production without large increase in marginal cost

What determines returns to scale?

By adopting the moving assembly line, Ford *chose* higher return to scale

- Better ability to scale up production without large increase in marginal cost

Many firm decisions affect their returns to scale

- **Production technique:** assembly lines, modularity
- **Organizational design:** managerial hierarchies, communication structure
- **Investment in scalable assets:** specialized machinery, automation processes, software

What determines returns to scale?

By adopting the moving assembly line, Ford *chose* higher return to scale

- Better ability to scale up production without large increase in marginal cost

Many firm decisions affect their returns to scale

- **Production technique:** assembly lines, modularity
- **Organizational design:** managerial hierarchies, communication structure
- **Investment in scalable assets:** specialized machinery, automation processes, software

Decisions by one firm affect others

- Ford's decision made the car a cheap input for the rest of the economy
- Cheaper inputs $\Longrightarrow$ incentives for *other* firms to scale up

What determines returns to scale?

By adopting the moving assembly line, Ford *chose* higher return to scale

- Better ability to scale up production without large increase in marginal cost

Many firm decisions affect their returns to scale

- **Production technique:** assembly lines, modularity
- **Organizational design:** managerial hierarchies, communication structure
- **Investment in scalable assets:** specialized machinery, automation processes, software

Decisions by one firm affect others

- Ford's decision made the car a cheap input for the rest of the economy
- Cheaper inputs $\implies$ incentives for *other* firms to scale up

What drives individual scalability decisions and how do they shape aggregate outcomes?

Approach and results

We construct a macroeconomic model with endogenous returns to scale

Approach and results

We construct a macroeconomic model with endogenous returns to scale

- Each firm can choose its returns to scale subject to a productivity cost
    - **Trade-off:** being effective at small vs. large scale

Approach and results

We construct a macroeconomic model with endogenous returns to scale

- Each firm can choose its returns to scale subject to a productivity cost
  - **Trade-off:** being effective at small vs. large scale

Key mechanism: Expansion incentives $\implies$ firms choose higher returns to scale.

Approach and results

We construct a macroeconomic model with endogenous returns to scale

- Each firm can choose its returns to scale subject to a productivity cost
  - **Trade-off:** being effective at small vs. large scale

Key mechanism: Expansion incentives $\Longrightarrow$ firms choose higher returns to scale.

- **Productivity:** top firms grow more $\Longrightarrow$ fatter upper tail of the size distribution

Approach and results

We construct a macroeconomic model with endogenous returns to scale

- Each firm can choose its returns to scale subject to a productivity cost
  - **Trade-off:** being effective at small vs. large scale

Key mechanism: Expansion incentives $\Longrightarrow$ firms choose higher returns to scale.

- **Productivity:** top firms grow more $\Longrightarrow$ fatter upper tail of the size distribution
- **Input prices:** propagation through the production network
  - Scalability begets scalability

Approach and results

We construct a macroeconomic model with endogenous returns to scale

- Each firm can choose its returns to scale subject to a productivity cost
  - **Trade-off:** being effective at small vs. large scale

Key mechanism: Expansion incentives $\implies$ firms choose higher returns to scale.

- **Productivity:** top firms grow more $\implies$ fatter upper tail of the size distribution
- **Input prices:** propagation through the production network
  - Scalability begets scalability

Aggregate implications

- Endogenous scalability allows top firms to grow larger $\implies$ higher GDP and GDP growth
- Distortions that affect the top firms are particularly harmful

Approach and results

We construct a macroeconomic model with endogenous returns to scale

- Each firm can choose its returns to scale subject to a productivity cost
  - **Trade-off:** being effective at small vs. large scale

Key mechanism: Expansion incentives $\implies$ firms choose higher returns to scale.

- **Productivity:** top firms grow more $\implies$ fatter upper tail of the size distribution
- **Input prices:** propagation through the production network
  - Scalability begets scalability

Aggregate implications

- Endogenous scalability allows top firms to grow larger $\implies$ higher GDP and GDP growth
- Distortions that affect the top firms are particularly harmful

We **calibrate** the model to the Spanish economy: large effects on GDP level and growth.

Literature review

- Classic work
  - Kuznets (1973), Chandler (1977, 1990)
- Endogenous scalability
  - Smirnyagin (2022), Lashkari et al. (2024), De Ridder (2024), Argente et al. (2025), Engbom et al. (2025), Gottlieb et al. (2025)
  - Contribution: tractable aggregation in general equilibrium, role of intermediate inputs, role of wedges
- Production function and RTS estimation
  - Hall (1990), Basu and Fernald (1997), De Loecker et al. (2020), Gao and Kehrig (2020), Demirer (2020), Ruzic and Ho (2023), Chiavari (2024), Chan et al. (2025).
  - Contribution: within-firm changes, impact of intermediate input prices, cross-country comparison
- Technique choice in production networks
  - Oberfield (2018), Acemoglu and Azar (2020), Kopytov et al. (2024, 2025)
  - Contribution: returns to scale as a technology choice

A model of endogenous returns to scale

Production

- Static model with competitive firms and representative household
- N sectors, each with a continuum of firms producing a homogenous good

Production

- Static model with competitive firms and representative household
- N sectors, each with a continuum of firms producing a homogenous good
- Firm l in sector i has a decreasing returns to scale Cobb-Douglas production function ▸ ς

$$F_i\left(L_{il}, X_{il}, \eta_{il}\right) = e^{\varepsilon_{il}} A_i\left(\eta_{il}\right) \zeta_{il}\left(\eta_{il}\right) \left(L_{il}^{1-\sum_{j=1}^{N} \alpha_{ij}} \prod_{j=1}^{N} X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}$$

  - $A_i\left(\eta_{il}\right)$ captures the cost of higher ret. to scale; $a_i\left(\eta_{il}\right) := \log A_i\left(\eta_{il}\right)$ strict. decreasing and concave
    - coordination and management costs, complications from more complex processes, etc.
  - Productivity draw $\varepsilon_{il} \sim$ iid $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$ is the only *source* of heterogeneity across firms within sector

Production

- Static model with competitive firms and representative household
- N sectors, each with a continuum of firms producing a homogenous good

- Firm l in sector i has a decreasing returns to scale Cobb-Douglas production function ▸ ζ

$$F_i\left(L_{il}, X_{il}, \eta_{il}\right) = e^{\varepsilon_{il}} A_i\left(\eta_{il}\right) \zeta_{il}\left(\eta_{il}\right) \left(L_{il}^{1-\sum_{j=1}^{N} \alpha_{ij}} \prod_{j=1}^{N} X_{ij,l}^{\alpha_{ij}}\right)^{\eta_{il}}$$

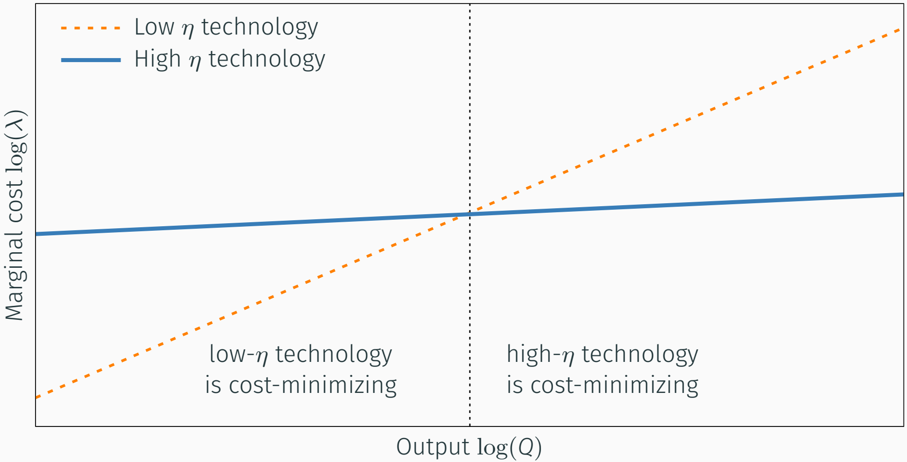
  - $A_i\left(\eta_{il}\right)$ captures the cost of higher ret. to scale; $a_i\left(\eta_{il}\right) := \log A_i\left(\eta_{il}\right)$ strict. decreasing and concave
    - coordination and management costs, complications from more complex processes, etc.
  - Productivity draw $\varepsilon_{il} \sim \text{iid } \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ is the only *source* of heterogeneity across firms within sector

- Firms maximize profits by jointly choosing inputs and returns to scale $\left(\eta_{il}\right)$

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

The Scalability Trade-off

High-η technologies are **better at large scale**; **low-η** technologies are **better at small scale**



Cost minimization problem

McKenzie (1959): A decreasing-returns technology can be interpreted as a constant-returns technology with a fixed entrepreneurial factor ($E_{il} = 1$).

$$Q_{il} = \underbrace{e^{\varepsilon_{il}} A_i\left(\eta_{il}\right)\zeta_i}_{\text{TFP}} \quad \times \quad \underbrace{V_{il}^{\eta_{il}}}_{\substack{\text{Variable inputs:} \\ \text{labor and interm.}}}$$

Cost minimization problem

McKenzie (1959): A decreasing-returns technology can be interpreted as a constant-returns technology with a fixed entrepreneurial factor ($E_{il} = 1$).

$$Q_{il} = \underbrace{e^{\varepsilon_{il}} A_i \left(\eta_{il}\right) \zeta_i}_{\text{TFP}} \quad \times \quad \underbrace{V_{il}^{\eta_{il}}}_{\substack{\text{Variable inputs:} \\ \text{labor and interm.}}} \quad \times \quad \underbrace{E_{il}^{1-\eta_{il}}}_{\substack{\text{Fixed entrep.} \\ \text{factor}}}$$

Cost minimization problem

McKenzie (1959): A decreasing-returns technology can be interpreted as a constant-returns technology with a fixed entrepreneurial factor ($E_{il} = 1$).

$$Q_{il} = \underbrace{e^{\varepsilon_{il}} A_i(\eta_{il})\,\zeta_i}_{\text{TFP}} \quad \times \quad \underbrace{V_{il}^{\eta_{il}}}_{\substack{\text{Variable inputs:}\\ \text{labor and interm.}}} \quad \times \quad \underbrace{E_{il}^{1-\eta_{il}}}_{\substack{\text{Fixed entrep.}\\ \text{factor}}}$$

Lemma

The firm's marginal cost of production is

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

where η_{il} governs the exposure to factor prices:

- $H_i = W^{1-\sum_{j=1}^{N} \alpha_{ij}} \prod_{j=1}^{N} P_j^{\alpha_{ij}}$ is price of the **variable input bundle** (Labor + Materials)
- Π_{il} is price of the **fixed factor** (Profits = Shadow cost of entrepreneur)

Optimal returns to scale

Lemma

The firm chooses its returns to scale $\eta_{il} \in (0,1)$ to minimize its marginal cost

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

Optimal returns to scale

Lemma

The firm chooses its returns to scale $\eta_{il} \in (0,1)$ to minimize its marginal cost

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

- Increasing η_{il} is costly and shifts input mix from fixed entrepreneurial factor to variable inputs
  - Cheaper variable inputs ($H_i \downarrow$) $\implies$ higher returns to scale ($\eta_{il} \uparrow$)
  - Any change pushing firm to be bigger (e.g., $\varepsilon_{il} \uparrow$ or $P_i \uparrow$) puts pressure on entrepreneurial factor which is in fixed supply $\implies$ firm relies less on it, i.e. η_{il} is higher

Optimal returns to scale

Lemma

The firm chooses its returns to scale $\eta_{il} \in (0,1)$ to minimize its marginal cost

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

- Increasing η_{il} is costly and shifts input mix from fixed entrepreneurial factor to variable inputs
  - Cheaper variable inputs ($H_i \downarrow$) $\implies$ higher returns to scale ($\eta_{il} \uparrow$)
  - Any change pushing firm to be bigger (e.g., $\varepsilon_{il} \uparrow$ or $P_i \uparrow$) puts pressure on entrepreneurial factor which is in fixed supply $\implies$ firm relies less on it, i.e. η_{il} is higher

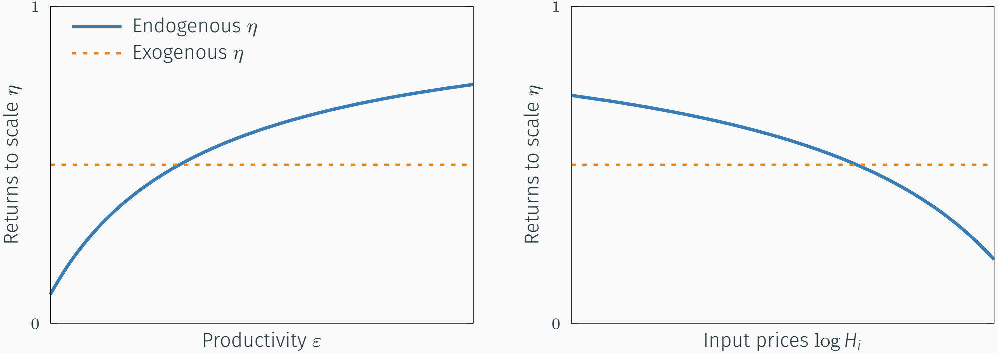
Lemma

Returns to scale η_{il} satisfy

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

(a) Returns to scale as a function of ε

(b) Returns to scale as a function of H_i



Impact on firm size: Double blessing of high ε_{il}

Lemma

Endogenous returns to scaleamplify the response of output

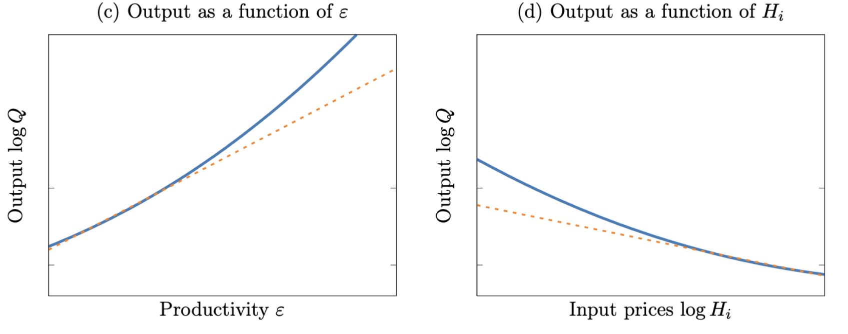
$$\frac{d\log Q_{il}}{d\varepsilon_{il}} = \underbrace{\frac{1}{1-\eta_{il}}}_{\text{Fixed } \eta \text{ effect}} + \underbrace{\frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\varepsilon_{il}}}_{\text{Superstar effect}} \quad \text{and} \quad \frac{d\log Q_{il}}{d\log H_i} = \underbrace{-\frac{\eta_{il}}{1-\eta_{il}}}_{\text{Fixed } \eta \text{ effect}} + \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log H_i}.$$

Impact on firm size: Double blessing of high ε_{il}

Lemma

Endogenous returns to scaleamplify the response of output

$$\frac{d\log Q_{il}}{d\varepsilon_{il}} = \underbrace{\frac{1}{1-\eta_{il}}}_{\text{Fixed } \eta \text{ effect}} + \underbrace{\frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\varepsilon_{il}}}_{\text{Superstar effect}} \quad \text{and} \quad \frac{d\log Q_{il}}{d\log H_i} = \underbrace{-\frac{\eta_{il}}{1-\eta_{il}}}_{\text{Fixed } \eta \text{ effect}} + \frac{1}{1-\eta_{il}}\frac{d\eta_{il}}{d\log H_i}.$$



Endogenous Returns to Scale Create Superstar Firms

Assumption for tractable aggregation

The cost function takes the form $a_i(\eta_{il}) = -\dfrac{\gamma_i}{1-\eta_{il}}$, where $\gamma_i > \sigma_i^2/2$. Let $\varphi_i := \sigma_i^2/(2\gamma_i) \in [0,1)$ denote the effective productivity dispersion in sector i.

Endogenous Returns to Scale Create Superstar Firms

Assumption for tractable aggregation

The cost function takes the form $a_i(\eta_{il}) = -\dfrac{\gamma_i}{1-\eta_{il}}$, where $\gamma_i > \sigma_i^2/2$. Let $\varphi_i := \sigma_i^2/(2\gamma_i) \in [0,1)$ denote the effective productivity dispersion in sector i.

- Endogenous ret. to scale leads to superstar firms and a thick tail of firm-size distribution

Endogenous Returns to Scale Create Superstar Firms

Assumption for tractable aggregation

The cost function takes the form $a_i(\eta_{il}) = -\frac{\gamma_i}{1-\eta_{il}}$, where $\gamma_i > \sigma_i^2/2$. Let $\varphi_i := \sigma_i^2/(2\gamma_i) \in [0,1)$ denote the effective productivity dispersion in sector i.

- Endogenous ret. to scale leads to superstar firms and a thick tail of firm-size distribution

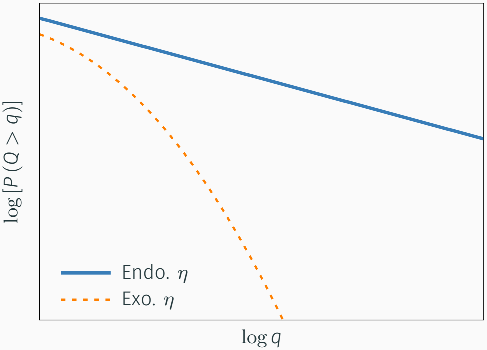
Proposition

With fixed returns to scale, firm size is log-normal. With endogenous returns to scale, the right tail becomes **Pareto**:

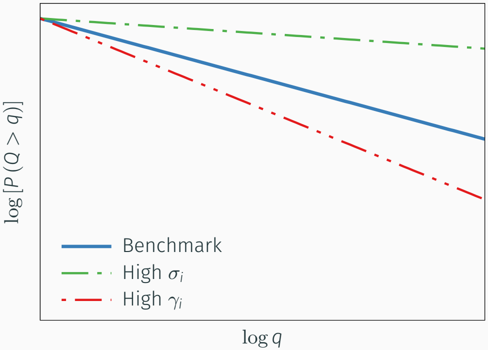
$$\log(\mathbb{P}(Q_{il} > q)) \sim -\frac{1}{\varphi_i} \log q, \text{ as } q \to \infty.$$

Pareto tail of firm-size distribution

(a) Impact of endogenous returns to scale



(b) Impact of σ_i and γ_i (endo. η)



Aggregation

Aggregation: Closing the model

1. **Free-entry condition:** Firms enter sector i until expected profits equal the entry cost $(\kappa_i W)$

$$\mathbb{E}_i \left[\Pi_{il} \left(\varepsilon_{il}, P, W \right) \right] = \kappa_i W$$

Aggregation: Closing the model

1. **Free-entry condition:** Firms enter sector i until expected profits equal the entry cost $(\kappa_i W)$

$$\mathbb{E}_i\left[\Pi_{il}\left(\varepsilon_{il}, P, W\right)\right] = \kappa_i W$$

2. **Representative household**
   - Supplies $\bar{L}$ units of labor inelastically
   - Cobb-Douglas preferences over sectoral goods

$$U\left(C\right) = \prod_{i=1}^{N}\left(\frac{C_i}{\beta_i}\right)^{\beta_i}$$

   - Budget constraint (profits are dissipated through entry costs)

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

Aggregation: Effective returns to scale

Definition: The effective returns to scale $\hat{\eta}_i$ in sector i is the sales-weighted average of η_{il}.

$$\hat{\eta}_i := \int_0^{M_i} \frac{P_i Q_{il}}{P_i Q_i} \eta_{il} \, dl$$

Aggregation: Effective returns to scale

Definition: The effective returns to scale $\hat{\eta}_i$ in sector i is the sales-weighted average of η_{il}.

$$\hat{\eta}_i := \int_0^{M_i} \frac{P_i Q_{il}}{P_i Q_i} \eta_{il} \, dl$$

Lemma

The returns to scale η_{il} of firm l in sector i can be expressed in terms of $\hat{\eta}_i$ and ε_{il}

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

- **Implication:** $\hat{\eta}_i$ is a sufficient statistic for the distribution of η_{il}
- **Selection effect:** $\hat{\eta}_i > \mathbb{E}[\eta_{il}]$. The effective scale is higher than the average because large firms are more scalable

Sectoral aggregation

- Free entry $\implies$ sector behaves like a **CRS** technology with endo. TFP $z_i(\hat{\eta}_i)$ and input shares $\hat{\eta}_i$

Proposition

The sectoral marginal cost of production is

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

where sectoral productivity $z_i(\hat{\eta}_i)$ decomposes into

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

- **Result:** Endogenous ret. to scale ($\varphi_i > 0$) boosts sectoral TFP through superstar effect

Proposition

1. In equilibrium, sectoral prices $P = (P_1, \ldots, P_N)$ equal marginal costs:

$$\log(P/W) = -\underbrace{\mathcal{L}(\hat{\eta})}_{\text{Network Multiplier}} \times \underbrace{z(\hat{\eta})}_{\text{Sectoral TFP}},$$

   where $\mathcal{L}(\hat{\eta}) = (I - \text{diag}(\hat{\eta})\,\alpha)^{-1}$ is the endogenous Leontief inverse matrix.

2. Equilibrium log GDP y is

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

   where $\omega_i = \frac{P_i Q_i}{PY} = \beta^\top \mathcal{L}(\hat{\eta})\,\mathbf{1}_i$ is the endogenous Domar weight of sector i.

General Equilibrium: Prices and GDP

Proposition

1. In equilibrium, sectoral prices $P = (P_1, \ldots, P_N)$ equal marginal costs:

$$\log (P/W) = - \underbrace{\mathcal{L}(\hat{\eta})}_{\text{Network Multiplier}} \times \underbrace{z(\hat{\eta})}_{\text{Sectoral TFP}},$$

   where $\mathcal{L}(\hat{\eta}) = (I - \mathrm{diag}(\hat{\eta})\,\alpha)^{-1}$ is the endogenous Leontief inverse matrix.

2. Equilibrium log GDP y is

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

   where $\omega_i = \frac{P_i Q_i}{PY} = \beta^{\top} \mathcal{L}(\hat{\eta})\,\mathbf{1}_i$ is the endogenous Domar weight of sector i.

Key insight: Returns to scale shape GDP through two channels

1. **Productivity** $z(\hat{\eta})$: Efficiency gains from superstar firms
2. **Network** $\omega(\hat{\eta})$: Higher $\hat{\eta}$ makes sectors more input-intensive $\implies$ higher Domar weights

Equilibrium returns to scale

Proposition: There exists a unique equilibrium and it is efficient $\implies \hat{\eta}$ maximizes GDP

Equilibrium returns to scale

Proposition: There exists a unique equilibrium and it is efficient $\Longrightarrow$ $\hat{\eta}$ maximizes GDP

Lemma

An increase in average productivity μ_j induces higher returns to scale in all downstream sectors:

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

where

1. $\Psi_i := (1-\varphi_i)\,\frac{d^2 a_i}{d\hat{\eta}_i^2} \leq 0$ captures how rigid $\hat{\eta}_i$ is
2. $\mathcal{K}_{ij} := \partial \log\left(H_i/W\right)/dz_j = -\left[\alpha\mathcal{L}\right]_{ij} \leq 0$ captures the impact of z_j on the price of i's input bundle

Equilibrium returns to scale

Proposition: There exists a unique equilibrium and it is efficient $\implies \hat{\eta}$ maximizes GDP

Lemma

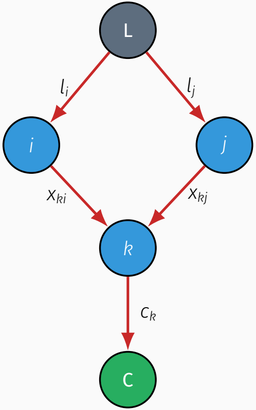
An increase in average productivity μ_j induces higher returns to scale in all downstream sectors:

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

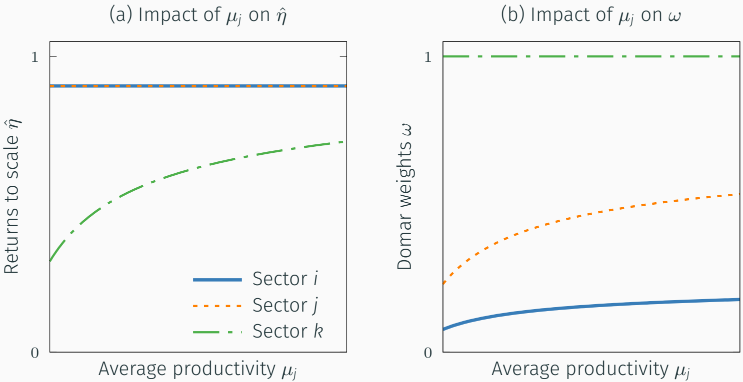
where

1. $\Psi_i := (1-\varphi_i)\frac{d^2 a_i}{d\hat{\eta}_i^2} \leq 0$ captures how rigid $\hat{\eta}_i$ is
2. $\mathcal{K}_{ij} := \partial \log(H_i/W)/dz_j = -[\alpha\mathcal{L}]_{ij} \leq 0$ captures the impact of z_j on the price of i's input bundle

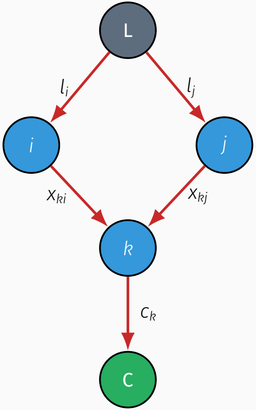
- Higher productivity μ_j lowers $P_j \implies$ all firms downstream of j increase $\hat{\eta}_i$

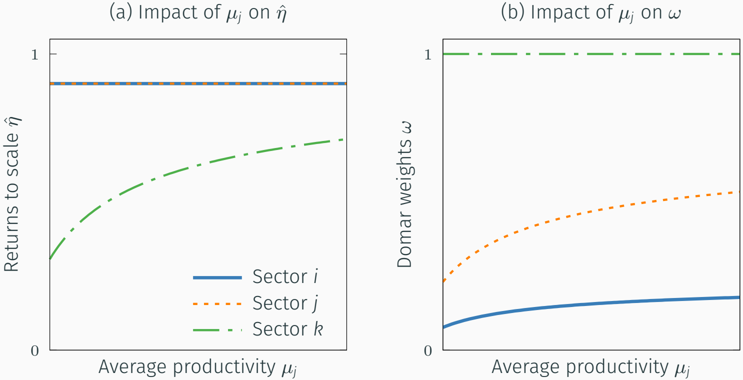


L
l_i
l_j
i
j
x_{ki}
x_{kj}
k
c_k
C



(a) Impact of μ_j on $\hat{\eta}$
1
0
Returns to scale $\hat{\eta}$
Average productivity μ_j
Sector i
Sector j
Sector k
(b) Impact of μ_j on ω
1
0
Domar weights ω
Average productivity μ_j





- Similar results for κ_j (lowers ret. to scale) and σ_j (increases ret. to scale)

Aggregate Implications

Endogenous returns to scale increase GDP

Define an alternative economy without endogenous returns to scale

Definition (Fixed returns-to-scale economy)

Fix all firms' returns to scale at the sectoral effective level ($\tilde{\eta}_{il} = \hat{\eta}_i$).

- By construction, this economy has the same sectoral Domar weights as the baseline.

Endogenous returns to scale increase GDP

Define an alternative economy without endogenous returns to scale

Definition (Fixed returns-to-scale economy)

Fix all firms' returns to scale at the sectoral effective level ($\tilde{\eta}_{il} = \hat{\eta}_i$).

- By construction, this economy has the same sectoral Domar weights as the baseline.

Proposition

Endogenous returns to scale increase the level of GDP

$$y - \tilde{y} = \sum_{i=1}^{N} \omega_i \frac{1}{2} (1 - \hat{\eta}_i) \log \left(\frac{1}{1 - \varphi_i} \right) > 0.$$

- With endogenous returns to scale, the most productive firms adopt the most scalable technology and grow disproportionately $\implies$ Resources reallocate to the most effective producers.

Proposition

The response of log GDP y to a shock $\Delta\mu_i$ is

$$\Delta y = \underbrace{\omega_i \Delta\mu_i}_{\substack{\text{Hulten's}\\ \text{theorem}}} + \underbrace{\frac{1}{2}\frac{d\omega_i}{d\mu_i}}_{\substack{\text{Endogenous}\\ \text{ret. to scale}}} (\Delta\mu_i)^2 + o\left((\Delta\mu_i)^2\right).$$

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

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

Response of GDP to shocks

Proposition

The response of log GDP y to a shock $\Delta\mu_i$ is

$$\Delta y = \underbrace{\omega_i \Delta\mu_i}_{\substack{\text{Hulten's}\\ \text{theorem}}} + \underbrace{\frac{1}{2}\frac{d\omega_i}{d\mu_i}}_{\substack{\text{Endogenous}\\ \text{ret. to scale}}} (\Delta\mu_i)^2 + o\left((\Delta\mu_i)^2\right).$$

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

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

- **Upstream/downstream propagation**
  - Higher $\mu_i \implies$ higher returns to scale downstream $\implies$ higher Domar weights upstream

Response of GDP to shocks

Proposition

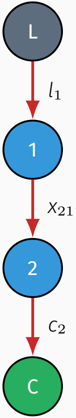
The response of log GDP y to a shock $\Delta\mu_i$ is

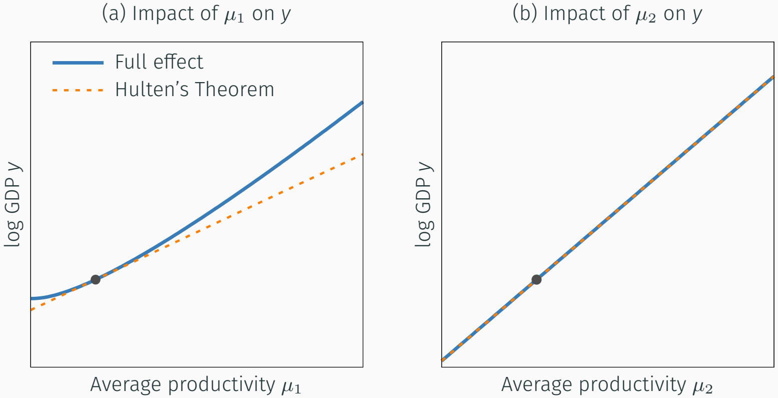
$$\Delta y = \underbrace{\omega_i \Delta\mu_i}_{\text{Hulten's theorem}} + \underbrace{\frac{1}{2}\frac{d\omega_i}{d\mu_i}}_{\text{Endogenous ret. to scale}} (\Delta\mu_i)^2 + o\left((\Delta\mu_i)^2\right).$$

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

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

- **Upstream/downstream propagation**
  - Higher $\mu_i \implies$ higher returns to scale downstream $\implies$ higher Domar weights upstream
- **Asymmetric response**
  - Magnifies the impact of positive shocks ($\Delta\mu_i > 0$)
  - Dampens the impact of negative shocks ($\Delta\mu_i < 0$)


L
l_1
1
x_{21}
2
c_2
C


(a) Impact of μ_1 on y
Full effect
Hulten's Theorem
log GDP y
Average productivity μ_1
(b) Impact of μ_2 on y
log GDP y
Average productivity μ_2

Implications for growth: Acceleration

One sector economy with constant TFP growth $d\mu/dt = g_\mu > 0$.

Implications for growth: Acceleration

One sector economy with constant TFP growth $d\mu/dt = g_\mu > 0$.

Lemma

As productivity rises, firms adopt more scalable technologies:

$$\frac{d\hat{\eta}}{dt} = -\Psi^{-1}\frac{\alpha}{1-\hat{\eta}\alpha}g_\mu > 0, \qquad \text{and} \qquad \lim_{t\to\infty}\hat{\eta} = 1.$$

- Productivity increases $\implies$ cheaper inputs $\implies$ higher $\hat{\eta}$

Implications for growth: Acceleration

One sector economy with constant TFP growth $d\mu/dt = g_\mu > 0$.

Lemma

As productivity rises, firms adopt more scalable technologies:

$$\frac{d\hat{\eta}}{dt} = -\Psi^{-1}\frac{\alpha}{1-\hat{\eta}\alpha}g_\mu > 0, \qquad \text{and} \qquad \lim_{t\to\infty}\hat{\eta} = 1.$$

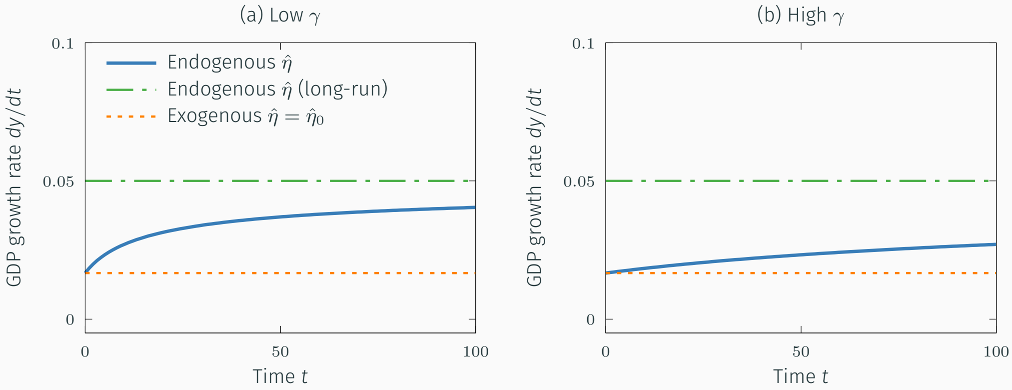
- Productivity increases $\implies$ cheaper inputs $\implies$ higher $\hat{\eta}$

Lemma

Endogenous scalability leads to strictly higher long-run growth:

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

- **Intuition:** Higher $\hat{\eta}$ $\implies$ Higher Domar weights $\implies$ Each increase in μ is more impactful


(a) Low γ
(b) High γ
0.1
0.05
0
GDP growth rate dy/dt
Endogenous $\hat{\eta}$
Endogenous $\hat{\eta}$ (long-run)
Exogenous $\hat{\eta} = \hat{\eta}_0$
0
50
100
Time t

Empirical evidence

Testing model predictions

Model: more productive firms should have higher returns to scale

$$\text{Cov}\left(\varepsilon_{il} + a_i\left(\eta_{il}\right), \eta_{il}\right) > 0.$$

This quantity also provides a measure of the strength of the endogenous scalability mechanism.

Testing model predictions

Model: more productive firms should have higher returns to scale

$$\text{Cov}\left(\varepsilon_{il} + a_i\left(\eta_{il}\right), \eta_{il}\right) > 0.$$

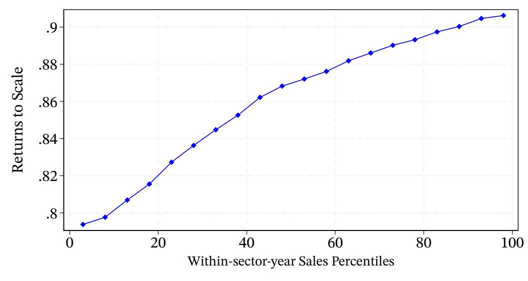
This quantity also provides a measure of the strength of the endogenous scalability mechanism.

Empirical strategy

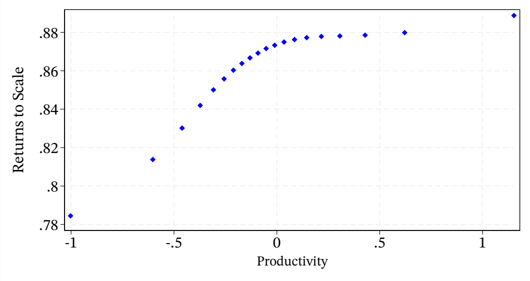
- **Data:** near-universe of Spanish firms (Orbis), 1995–2019 (9,754,405 firm-year observations)
- **Methodology:**
  - Group firms into size deciles within each sector-year
  - Estimate production functions (CD with K, L, M) for each decile using Blundell-Bond (2000).
  - Recover returns to scale as sum of output elasticities
- Productivity is measured using a Törnqvist input-quantity index to handle heterogeneous technologies

More productive firms have higher returns to scale

Model prediction: more productive firms should have higher returns to scale



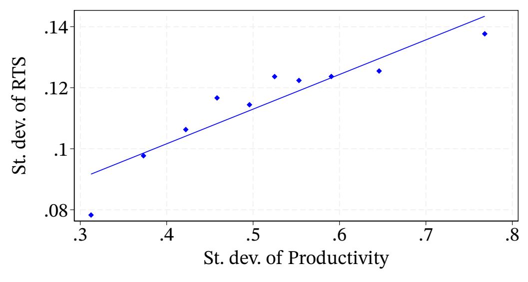
(a) Returns to scale and firm size



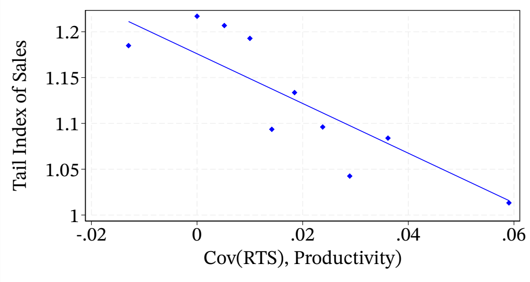
(b) Returns to scale and productivity

Model predictions

1. Sectors with more dispersed productivity should have more dispersed ret. to scale
2. Sectors with stronger endo. ret. to scale mechanism should have thicker tail of firm-size dist.
   - Covariance between ret. to scale and productivity as proxy



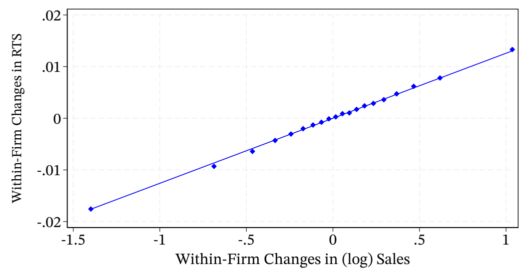
(a) Dispersion in returns to scale and productivity



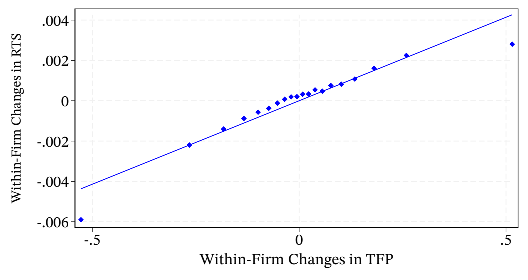
(b) Tail indices of sales and endogenous scalability

Within-firm evidence: Firms increase η as they grow

Model prediction: As firms become more productive, they increase their returns to scale



(a) Returns to scale and sales



(b) Returns to scale and productivity

Note: Regression absorbing firm and sector-year fixed effects

Testing the cost channel: Tariff shocks

Model prediction: Higher input prices force firms to reduce returns to scale:

Testing the cost channel: Tariff shocks

Model prediction: Higher input prices force firms to reduce returns to scale:

Empirical approach:

- Local projections using tariff-induced input cost shocks (T_{it}) from Teti (2024):

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

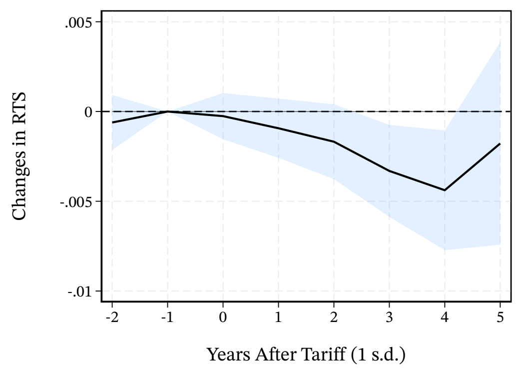
Testing the cost channel: Tariff shocks

Model prediction: Higher input prices force firms to reduce returns to scale:

Empirical approach:

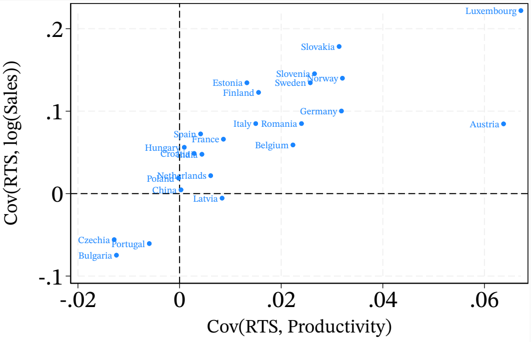
- Local projections using tariff-induced input cost shocks (T_{it}) from Teti (2024):

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

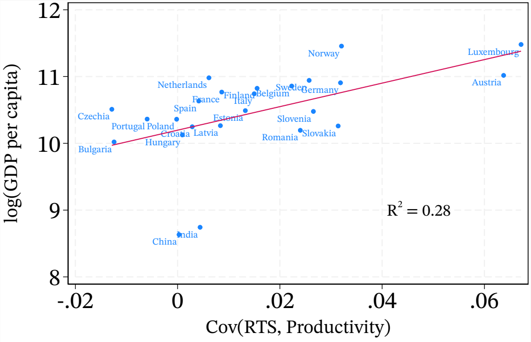


Endogenous scalability and development

- We replicate the estimation for manufacturing firms in **24 countries**.



(a) Endogenous scalability across countries



(b) Economic development and endogenous scalability

- **Left:** the mechanism is visible globally ($\text{Cov}(\eta, \text{TFP}) > 0$)
- **Right:** Richer countries have stronger endogenous scalability

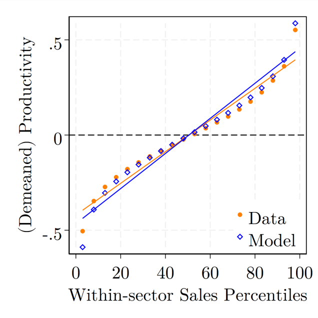
Calibration

Calibration

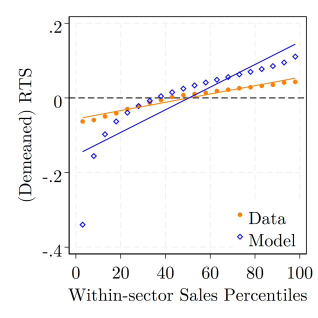
- We calibrate the model to the Spanish economy (62 sectors)
- Some parameters have direct empirical counterparts
  - Consumption shares β and supply chain structure α
- Left to choose: productivity parameters μ and σ; cost function parameter γ; entry cost κ
  - σ and γ govern within-sector heterogeneity
    - Target sectoral interquartile range in log profits and RTS (Bloom et al., 2018)
  - No need to pick μ and κ; they always can be chosen to match $\hat{\eta}$

▸ Fit

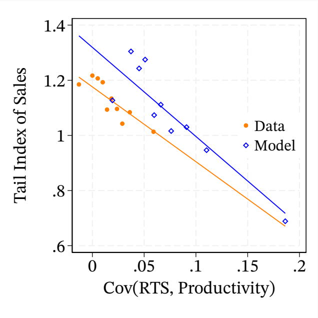
Untargeted moments: Productivity, size, and returns to scale



(a) Productivity and sales



(b) Ret. to scale and sales



(c) Ret. to scale and productivity

Implications for GDP

How much does endogenous scalability matter for the **level of GDP**? $y - \tilde{y} \approx$**10%**

- Better allocation of resources to most productive firms

Implications for GDP

How much does endogenous scalability matter for the **level of GDP**? $y - \tilde{y} \approx$**10%**

- Better allocation of resources to most productive firms

Impact of endogenous ret. to scale on the **growth rate of GDP**

- Constant productivity growth in all sectors $(\mu_i(t) = \mu_i(0) + 0.01 \times t\ )$
- Consider three economies
  - **Baseline**: firms are free to adjust returns to scale, so they pick $\eta_{il}(\mu(t))$
  - **Dispersed RTS**: firms keep their initial RTS $\eta_{il}(\mu(0))$
  - **Fixed RTS**: firms' returns to scale is set to the initial sectoral average $\hat{\eta}_i(\mu(0))$

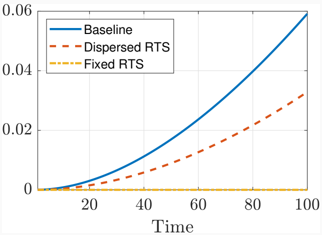
Implications for GDP

How much does endogenous scalability matter for the **level of GDP**? $y - \tilde{y}$ ≈**10%**

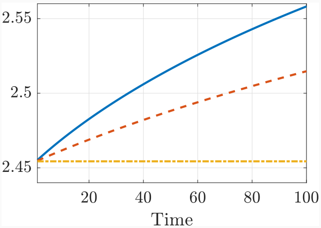
- Better allocation of resources to most productive firms

Impact of endogenous ret. to scale on the **growth rate of GDP**

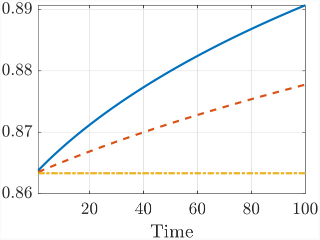
- Constant productivity growth in all sectors ($\mu_i(t) = \mu_i(0) + 0.01 \times t$)
- Consider three economies
  - **Baseline**: firms are free to adjust returns to scale, so they pick $\eta_{il}(\mu(t))$
  - **Dispersed RTS**: firms keep their initial RTS $\eta_{il}(\mu(0))$
  - **Fixed RTS**: firms' returns to scale is set to the initial sectoral average $\hat{\eta}_i(\mu(0))$



(a) Log GDP relative to fixed RTS



(b) GDP growth rate [%]



(c) Average $\hat{\eta}_i$

Wedges

- So far we have considered an efficient economy. What if we introduce wedges?
- Measure sales wedges a la Hsieh and Klenow (2009): $\frac{1}{1-\tau_i^S} = \frac{\text{MRPL}}{W}$
    - Fit the level + size-dependent component: $\log\left(1-\tau_{il}^S\right) = \log\left(1-\tau_i^S\right) - b_i\left(\varepsilon_{il} - \mu_i\right),$
    - $b_i > 0 \implies$ Large firms face higher distortions.

Wedges

- So far we have considered an efficient economy. What if we introduce wedges?
- Measure sales wedges a la Hsieh and Klenow (2009): $\frac{1}{1-\tau_i^S} = \frac{\mathrm{MRPL}}{W}$
  - Fit the level + size-dependent component: $\log\left(1-\tau_{il}^S\right) = \log\left(1-\tau_i^S\right) - b_i\left(\varepsilon_{il} - \mu_i\right)$,
  - $b_i > 0 \implies$ Large firms face higher distortions.

Table 1: Returns to scale and GDP when wedges are removed

	Size-dependent wedges		Flat wedges	
	Δ Ret. to scale	Δ GDP	Δ Ret. to scale	Δ GDP
Baseline economy	0.067	167%	0.020	62%
Fixed ret. to scale	0	70%	0	58%

- Gains are **>2x larger** in the Baseline vs. Fixed model (167% vs 70%).
- Wedges that affect the top firms are responsible for most of the action

More

Conclusion

Conclusion

Main contributions

- Tractable multisector model with endogenous returns to scale
    - Input-output linkages play a crucial role in driving mechanisms
- Matches key patterns in Spanish +cross-country data
- Substantial quantitative effect on level and growth rate of GDP

More results in the paper

- Comparative static with respect to key parameters
- Analytical expression for growth rate along transition path
- Full-fledged model with wedges

Future work

- Role of capital
- Interaction of returns to scale with market power
- Individual margins that affect returns to scale (microfoundation for A_i)

Expression for $\zeta(\alpha_i)$

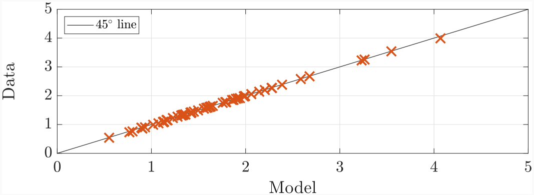
The function $\zeta_{il}(\alpha_i)$ is

$$\zeta_{il}(\eta) := \left[\left(\left(1 - \sum_j \alpha_{ij} \right) \eta \right)^{\eta\left(1-\sum_j \alpha_{ij}\right)} \prod_j (\eta\alpha_{ij})^{\eta\alpha_{ij}} (1-\eta)^{1-\eta} \right]^{-1}$$

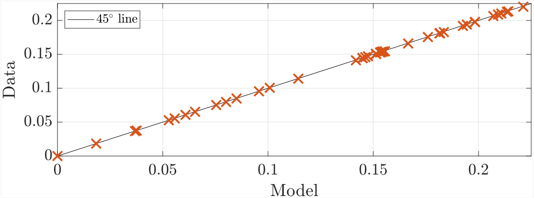
This functional form allows for a simple expression for the unit cost K

Back

Fit of the calibrated model



(a) IQR of log profits



(b) IQR of returns to scale

◀ Back

Impact of wedges on calibrated economy

Table 2: Returns to scale and GDP when wedges are removed

	Size-dependent wedges		Flat wedges	
	Δ Ret. to scale	Δ GDP	Δ Ret. to scale	Δ GDP
Baseline economy	0.067	167%	0.020	62%
Dispersed ret. to scale	0.046	138%	0.010	60%
Fixed ret. to scale	0	70%	0	58%

Back