Returns to Scale and Productivity in the Macroeconomy

Joel Kariel[*] Anthony Savagar[†]

October 17, 2022

Find latest draft here.

Abstract

We develop a model of heterogeneous firms with endogenous returns to scale in order to study the relationship between returns to scale and productivity. We show that rising returns to scale in variable production, as opposed to fixed costs, cause returns to scale to increase whilst productivity declines. This occurs because improving returns in variable production weakens firm selection, which allows low-productivity firms to survive. In our empirical analysis, we estimate returns to scale and productivity for a panel of U.K. firms. We find that across firms productivity and returns to scale are negatively related, and in aggregate returns to scale have increased whilst productivity has stagnated. Broadly, our results help to explain the puzzle of rising returns to scale but weak productivity in several advanced economies.

JEL: E32, E23, D21, D43, L13.

Keywords: Returns to Scale, Productivity, Market Structures, Firm Dynamics.

[*] Oxford University and University of Kent, joel.kariel@economics.ox.ac.uk

[†] University of Kent, a.savagar@kent.ac.uk.

This research is funded under ESRC project reference ES/V003364/1.

Disclaimer: *This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.*

We thank attendees at the Kent research seminar and the Bristol Macro seminar. We also thank participants from IFN Stockholm, EEA-ESEM 2022, IAAE 2022, CEF 2022, RES 2022, SNDE 2022, AMEF 2022, SES 2022, MMF 2022, the Kent Firm Dynamics Workshop 2021, and the Exeter Workshop 2022 for their helpful comments. Thanks to Omar Licandro, Julian Neira, Tom Schmitz, Petr Sedláček, Danial Lashkari, Steve Bond and Kunal Sangani for detailed feedback and discussions.

Motivation

Returns to scale are an important factor in recent market power discussions. Many explanations of rising market power, such as intangible assets, superstar firms, IT-investment, and regulatory barriers can be interpreted through returns to scale, either by affecting fixed costs or variable costs. Our paper studies whether returns to scale caused by variable cost changes or fixed cost changes affect productivity differently. We find that changing returns to scale driven by variable cost changes is more consistent with the returns to scale and productivity relationships we observe in the data.

Findings

What is the relationship between returns to scale and productivity? Do the sources of returns to scale (marginal cost versus fixed cost changes) matter for aggregate outcomes? We address these questions by developing a heterogeneous firm model with endogenous returns to scale, and testing the model predictions against estimated returns to scale and productivity for a panel of U.K. firms.

Our theoretical findings show that returns to scale and productivity are negatively related if endogenously higher returns to scale are driven by variable costs, whereas there is no relationship between endogenous returns to scale and productivity through changing fixed costs. Given this theoretical insight, our data analysis tests the relationship between productivity and returns to scale. Our empirical findings show that the two variables are negatively related in the data which implies changing variable costs are the main driver of returns to scale. If changing fixed costs were responsible, then we would not observe the productivity relationship that we observe. In addition to documenting a negative relationship between returns to scale and productivity, we show that overall returns to scale in the UK have increased whilst productivity has declined. The result of rising returns to scale supporting low-productivity firm survival is consistent with evidence that shows a major cause of low productivity in the UK is survival of small unproductive firms (Barnett, Batten, Chiu, Franklin, and Sebastia-Barriel 2014; Riley, Rosazza-Bondibene, and Young 2015).

Methodology

We develop a neoclassical growth model with heterogeneous firms, endogenous returns to scale, and monopolistic competition based on Hopenhayn and Rogerson (1993), Restuccia and Rogerson (2008), Barseghyan and DiCecio (2011), and Barseghyan and DiCecio (2016).[1] We extend this framework to distinguish different types of returns to scale – those internal to the firm versus external returns from aggregation – and focus on the drivers of endogenous returns to scale. Additionally we emphasize analytical results from a Pareto productivity distribution. We focus on steady-state outcomes and our main theoretical exercise compares the outcomes for endogenously determined returns to scale and productivity when we change variable costs and fixed costs. This analysis highlights that whether changing variable costs or changing fixed costs are responsible for returns to scale will lead to different outcomes for productivity. We test this relationship in the data, and the outcome we observe in the data is consistent with variable costs being the driver of returns to scale across firms.

Why does raising returns to scale in variable production cause overall returns to scale to rise whilst productivity falls? Rising returns in variable production raises overall returns to scale directly, but reduces average productivity by allowing less productive firms to compete. The mechanism rests on reducing wages which makes it easier for low-productivity firms to pay labour-denominated fixed costs, consequently more low-productivity firms survive which decreases average productivity.

To measure returns to scale and productivity, we estimate firm-level production functions using a methodology developed by Ackerberg, Caves, and Frazer (2015). This methodology overcomes endogeneity issues caused by unobserved productivity at the firm level. The sum of coefficients on factors of production from these regressions yield returns to scale and the residuals yield productivity. Our main dataset is the Annual Business Survey (ABS) which is a representative sample of roughly 60,000 firms in the U.K. each year 1998-2014 covering two-thirds of aggregate value-added.

[1] Neoclassical growth models with endogenous industry structures have been developed by various authors such as Atkeson and P. J. Kehoe (2005), Collard and Licandro (2021), and Asturias, Hur, T. J. Kehoe, and Ruhl (2022). Licandro (2022) provides a detailed survey.

A well-known application of the dataset is Aghion, Bloom, Blundell, Griffith, and Howitt (2005).

Literature

Recent research on changing market structures – rising markups, rising profit shares, declining business dynamism – suggests *technological factors* in production (Van Reenen 2018; Autor, Dorn, Katz, Patterson, and Van Reenen 2017; Lashkari, Bauer, and Boussard 2019; Bessen 2020; De Ridder 2019; Aghion, Bergeaud, Boppart, Klenow, and Li 2019; Foster, J. C. Haltiwanger, and Tuttle 2022) or *behavioural changes* (Gutierrez, Jones, and Philippon 2019; Barkai 2020; De Loecker, Eeckhout, and Mongey 2021) in the optimizing behaviour of firms may have caused changing market structures. There is also literature that attempts to understand the implications for other macroeconomic variables and welfare (Edmond, Midrigan, and Xu 2021; De Loecker, Eeckhout, and Mongey 2021; Gutiérrez, Jones, and Philippon 2021; Eggertsson, Robbins, and Wold 2021). These papers develop dynamic general equilibrium models with firm entry and imperfect competition. Baqaee, Farhi, and Sangani (2021) focus directly on returns to scale to show how reallocation of output towards larger firms can be welfare enhancing if they benefit from returns to scale.

Many of the factors that other researchers have investigated implicitly affect returns to scale, either through marginal costs or fixed costs. Our paper focuses on this distinction in the drivers of returns to scale and whether this matters for the impact of returns to scale on productivity. Kim (2021) studies the importance of the slope of the marginal cost curve for international business cycles.

Several recent papers provide estimates of returns to scale in the U.S. economy. Gao and Kehrig (2021) estimate returns to scale of 0.96 overall in U.S. manufacturing firms, and returns to scale across 4-digit industries, ranging from 0.86 to 1.3. Ruzic and Ho (2019) estimate returns to scale at the industry level. They find a decline in returns to scale from 1.2 in 1982 to 0.96 in 2007. Lashkari, Bauer, and Boussard (2019) find returns to scale ranging from 0.75 to 1.06, with smaller firms obtaining larger scale economies. The most recent estimates of returns to scale in the U.K. economy are Oulton (1996), Harris and Lau (1998), and Girma and Görg (2002). These studies document constant or slightly decreasing returns to scale for manufacturing firms.

1 Model

The model setup follows Barseghyan and DiCecio (2016) with the addition of external scale economies and Pareto distributed productivity. The household side of the model follows a standard neoclassical growth setup. The firm side of the economy follows a two-stage setup. First, a firm decides whether to pay a fixed entry cost based on the expected profits they receive from optimal production decisions. Second, given a firm has entered, it makes optimal production decisions. It receives a productivity draw at which point it decides to produce zero or a positive amount. This decision is based on whether producing output will generate enough profit to cover a fixed, period-by-period, overhead cost. At the end of the period all firms die exogenously.

Heterogeneous productivity leads to heterogeneous firm output, but the fixed cost is the same for all firms. This leads to a different returns to scale across firms. In addition to returns to scale from the fixed cost, firms also have decreasing returns to variable inputs (upward sloping marginal cost curves). If a firm has low productivity, it will be small and the fixed cost will dominate the decreasing returns in variable production leading to increasing returns to scale. If the firm has high productivity, it will be large and the decreasing returns in variable inputs will offset the increasing returns from the overhead, so overall the firm has decreasing returns.

1.1 Households

A representative household maximizes lifetime utility subject to a budget constraint

$$\max_{\{C_t,K_{t+1}\}_{t=0}^{\infty}} \quad \sum_{t=0}^{\infty}\beta^t \frac{C_t^{1-\sigma}-1}{1-\sigma}, \quad \beta\in(0,1),$$

$$\text{s.t.} \quad C_t + I_t = r_t K_t + w_t L^s + \Pi_t + T_t \tag{1}$$

$$I_t = K_{t+1} - (1-\delta)K_t. \tag{2}$$

Households own all firms in the economy and receive profit Π_t. T_t is a lump-sum transfer from the government which will equal to the entry fees paid by firms. Households supply a fixed amount of labour that is not time-varying, we normalize this to one:

$$L^s = 1. \tag{3}$$

Households own the capital stock and rent it to firms at a rental rate r_t, hence the capital investment decision is part of the household problem. The household optimization problem satisfies the following condition

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta(r_{t+1} + (1-\delta)). \tag{4}$$

plus a transversality condition and the resource constraint.

1.2 Firms

The variable $\imath \in (0, N_t)$ indexes active firms, whereas $\jmath \in (0,1)$ indexes all entrants, both active and inactive. There is a cut-off J_t on the unit interval below which firms are inactive and above which firms are active. Therefore $\imath$ is a subset of entrants corresponding to $\jmath \in (J_t, 1)$ and those in $(0, J_t)$ are inactive.

1.2.1 Final goods producer

The final goods aggregator is

$$Y_t = N_t^{1+\epsilon}\left[\frac{1}{N_t}\int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}} d\imath\right]^{\mu}. \tag{5}$$

There are N_t intermediate producers on the interval $\imath \in (0, N_t)$ and $\mu \in (1, 1 + \nu \min(\alpha, 1-\alpha))$ captures product substitutability. It will turn out that μ is the price markup that monopolistically competitive firms charge. The $1+\epsilon$ term captures external returns to scale, also known as agglomeration economies or love of variety when introduced in a consumption aggregator. If $1+\epsilon = \mu$ the aggregator is a standard CES with implicit love of variety. If $1+\epsilon = 1$ then there is no love of variety.

The maximization problem of the final goods producer is

$$\Pi_t^F = \max_{y_t(\imath)} \quad Y_t - \int_0^{N_t} p_t(\imath) y_t(\imath) d\imath \tag{6}$$

$$\text{s.t.} \quad Y_t = N_t^{1+\epsilon}\left[\frac{1}{N_t}\int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}} d\imath\right]^{\mu} \tag{7}$$

The firm is infinitesimal so firm level output does not affect Y_t. The first-order condition with respect to $y_t(\imath)$ gives the inverse-demand for a firm

$$p_t(\imath) = N_t^{\frac{1+\epsilon-\mu}{\mu}} \left(\frac{y_t(\imath)}{Y_t}\right)^{\frac{1-\mu}{\mu}}. \tag{8}$$

1.2.2 Intermediate goods producer

The timeline for the intermediate goods producer is that the firm pays cost κ to enter. It receives a productivity draw $\jmath \in (0,1)$ then decides whether to produce which incurs a fixed overhead cost. If the firm does not produce it remains inactive which we refer to as endogenous exit. All firms, active and inactive, exit exogenously at the end of one period.

The production function for firm with productivity $\jmath$ is given by

$$y(\jmath) = A(\jmath)^{1-\nu}\left[k(\jmath)^{\alpha}\ell(\jmath)^{1-\alpha}\right]^{\nu}, \quad \nu \in [0,1]. \tag{9}$$

The parameter $\nu \in [0,1]$ captures diminishing returns in variable production (i.e. upward sloping marginal cost curve). The labour employed to produce output is:

$$\ell_t(\jmath) = \ell_t^{\text{tot}}(\jmath) - \phi, \tag{10}$$

where $\ell_t^{\text{tot}}(\jmath)$ represents the total labour employed by the firm, and ϕ is a labour-denominated fixed overhead cost. Both ϕ and ν determine returns to scale at the firm level. The productivity term $A(\jmath)$ is the inverse of the CDF of the productivity distribution. In the Appendix we illustrate this using the Pareto distribution.

The firm solves the following profits maximization problem:

$$\max_{k_t(\jmath),\ell_t(\jmath),y_t(\jmath),p_t(\jmath)} p_t(\jmath)y_t(\jmath) - r_t k_t(\jmath) - w_t(\ell_t(\jmath)+\phi) \tag{11}$$

subject to the production function (9) and inverse demand function (8). The optimality conditions imply

$$\frac{r_t}{p_t(\jmath)} = \frac{\nu}{\mu}\alpha\frac{y_t(\jmath)}{k_t(\jmath)} \tag{12}$$

$$\frac{w_t}{p_t(\jmath)} = \frac{\nu}{\mu}(1-\alpha)\frac{y_t(\jmath)}{\ell_t(\jmath)}. \tag{13}$$

1.2.3 Ratio of firm size

The inverse demand condition and factor price equilibrium conditions imply that for any two firms $\imath, \jmath$ revenue and input choice are proportional to scaled productivity:

$$\frac{p_t(\jmath)y_t(\jmath)}{p_t(\imath)y_t(\imath)} = \frac{k_t(\jmath)}{k_t(\imath)} = \frac{\ell_t(\jmath)}{\ell_t(\imath)} = \frac{a(\jmath)}{a(\imath)}, \quad \forall \imath, \jmath, \quad \text{where} \quad a(\jmath) \equiv A_t(\jmath)^{\frac{1-\nu}{\mu-\nu}}. \tag{14}$$

1.2.4 Zero-profit firm

We assume there is a threshold productivity level $J_t \in (0,1)$ characterised by zero profits. If a firm receives a productivity draw below the threshold productivity level they would make negative profits from production. Consequently, they prefer to produce zero and make zero profits. Therefore we define profits and characterise the threshold productivity as follows:

$$\pi_t(\jmath) = p_t(\jmath)y_t(\jmath) - r_t k_t(\jmath) - w_t(\ell_t(\jmath) + \phi) \tag{15}$$

$$\pi_t(J_t) = 0. \tag{16}$$

A helpful reduced-form expression for profits combines the profit condition with equilibrium factor prices, with the zero-profit condition and with the ratio of revenues to scaled productivity (see appendix for details):

$$\pi_t(\jmath) = \phi w_t \left(\frac{a(\jmath)}{a(J_t)} - 1 \right). \tag{17}$$

Profits are rising in fixed costs and relative productivity of the firm.

1.2.5 Free Entry

All firms die after one period. A firm only produces if it makes positive profits, hence firm value is given by

$$v_t(\jmath) = \max\{\pi_t(\jmath), 0\}. \tag{18}$$

We assume a free entry condition which implies that the expected value from entering equals to the entry cost κ:

$$\mathbb{E}[v_t(\jmath)] = \kappa. \tag{19}$$

Combining (18) and (19) with our reduced-form profit expression (17) yields

$$\phi w_t (1 - J_t) \left[\frac{\bar{a}(J_t)}{a(J_t)} - 1 \right] = \kappa. \tag{20}$$

We use bar notation to represent the mean value of a function.[2] The mean of scaled productivity $a(\jmath)$ over the interval $(J_t, 1)$ is

$$\bar{a}(J_t) = \frac{1}{1 - J_t} \int_{J_t}^{1} a(\jmath) d\jmath. \tag{21}$$

1.3 Entry

Operating firms N_t are the subset of firms who decide to produce once receiving their productivity draw. Entrants E_t are all firms who pay the entry cost.

$$N_t = \int_0^{N_t} d\imath = E_t \int_{J_t}^{1} d\jmath = E_t(1 - J_t). \tag{22}$$

We can interpret the productivity cut-off J_t as the fraction of entering firms who choose to produce 0 after receiving their productivity draw.

1.4 Aggregation

To get aggregate output and aggregate inputs, note that the integral over the index of operating firms $(0, N_t)$ is equivalent to entering firms E_t constrained over the region of operation $(J_t, 1)$.

1.4.1 Aggregate Factor Inputs

Aggregate labour is comprised of production labour and non-production labour

$$K_t = \int_0^{N_t} k_t(\imath) d\imath = E_t \int_{J_t}^{1} k_t(\jmath) d\jmath \tag{23}$$

$$L_t = \int_0^{N_t} [\ell_t(\imath) + \phi] d\imath = E_t \int_{J_t}^{1} [\ell_t(\jmath) + \phi] d\jmath. \tag{24}$$

[2] The mean value of f on $[a, b]$ is defined as

$$\bar{f}(x) \equiv f(\bar{x}) = \frac{1}{b - a} \int_a^b f(x) dx.$$

We define u_t as the fraction of aggregate labour that goes to production

$$u_t \equiv \frac{E_t \int_{J_t}^{1} \ell(\jmath) d\jmath}{L_t} \tag{25}$$

$$1 - u_t = \frac{E_t(1 - J_t)\phi}{L_t} = \frac{N_t \phi}{L_t}. \tag{26}$$

1.4.2 Aggregate Output

In the appendix we present a derivation of aggregate output.

$$Y_t = N_t^{1+\epsilon-\nu} \bar{a}(J_t)^{\mu-\nu} \left[K_t^{\alpha} (u_t L_t)^{1-\alpha} \right]^{\nu}. \tag{27}$$

Using the expression for labour used in non-production, we can remove $N_t = \frac{1-u_t}{\phi} L_t$, which yields aggregate output as a Cobb-Douglas function of *aggregate* inputs

$$Y_t = \mathrm{TFP}_t K_t^{\alpha\nu} L_t^{1+\epsilon-\alpha\nu} \quad \text{where} \quad \mathrm{TFP}_t \equiv \left(\frac{1 - u_t}{\phi} \right)^{1+\epsilon-\nu} u_t^{(1-\alpha)\nu} \bar{a}(J_t)^{\mu-\nu}. \tag{28}$$

The pre-multiplying term represents aggregate total factor productivity. That is, it captures changes in aggregate output that are not accounted for by changes in aggregate inputs. TFP is not the Solow residual because the exponents of aggregate capital and labour do not correspond to aggregate factor shares.

1.4.3 Aggregate Factor Market Equilibrium

The wage, rental rate on capital and zero profit condition are

$$r_t = \alpha \frac{\nu}{\mu} \frac{Y_t}{K_t} \tag{29}$$

$$w_t = (1 - \alpha) \frac{\nu}{\mu} \frac{Y_t}{u_t L_t} \tag{30}$$

$$\frac{w_t (1 - u_t) L_t}{Y_t} = \left(1 - \frac{\nu}{\mu} \right) \frac{a(J_t)}{\bar{a}(J_t)} \tag{31}$$

1.5 Government Budget Constraint and Resource Constraints

The resource constraint is

$$Y_t = C_t + I_t. \tag{32}$$

The government rebates entry fees to households. The government budget constraint equates taxes to government expenditure

$$T_t = E_t \kappa. \tag{33}$$

Profits and labour markets clear

$$\Pi_t = \Pi_t^F \tag{34}$$

$$L_t = L^s. \tag{35}$$

1.6 Equilibrium Definition

An equilibrium is a sequence of prices $\{r_t, w_t\}_{t=0}^{\infty}$; firm capital and labour demands $\{\ell_t(\jmath), k_t(\jmath)\}_{t=0}^{\infty}$; firms' operating decisions, measures of entry and operation $\{E_t, N_t\}_{t=0}^{\infty}$; consumption and capital $\{C_t, K_{t+1}\}_{t=0}^{\infty}$, such that

1. households choose C and K optimally by solving problem (1);
2. firms compete under monopolistic competition and decide optimally whether to produce or remain dormant, and factor demands satisfy (11);
3. the free entry condition holds (19);
4. markets clear for aggregate labour (24), aggregate capital (23), goods market (32), labour market (35) and aggregate profits (34);
5. the government budget constraint is satisfied (33).

2 Model Analysis

In this section we apply the assumption of a Pareto productivity distribution $A(\jmath)$ in order to obtain analytical expressions for productivity and returns to scale in terms of parameters and the productivity cut-off J_t. We analyse the steady-state of the model and discuss the implications of changing parameters on our outcomes of interest.

2.1 Average Productivity with Pareto Distribution

If we assume a Pareto distribution, scaled productivity is given by

$$a(\jmath) = A(\jmath)^{\Gamma} = \frac{1}{(1-\jmath)^{\frac{\Gamma}{\vartheta}}}. \tag{36}$$

The term $\Gamma \equiv \frac{1-\nu}{\mu-\nu} \in (0,1)$ is the productivity scaling exponent and ϑ is the Pareto tail parameter. Average productivity is a linear function of the productivity level of the cut-off firm:

$$\frac{a(J_t)}{\bar{a}(J_t)} = 1 - \frac{\Gamma}{\vartheta}. \tag{37}$$

We can express average productivity as:

$$\bar{A}(J_t) = \frac{\vartheta}{\vartheta-1}(1-J_t)^{-\frac{1}{\vartheta}} \tag{38}$$

Changes to the productivity cut-off J_t determine average productivity.

2.2 Returns to Scale

2.2.1 External Returns to Scale

Changes in aggregate output occur through three channels: changes in TFP, changes in aggregate capital and changes in aggregate labour. The total derivative of aggregate output is:

$$d\ln Y_t = \frac{\partial \ln Y_t}{\partial \ln TFP_t} d\ln TFP_t + \frac{\partial \ln Y_t}{\partial \ln K_t} d\ln K_t + \frac{\partial \ln Y_t}{\partial \ln L_t} d\ln L_t \tag{39}$$

From equation (28), the degree of returns to scale in the aggregate economy is:

$$\text{External RTS} \equiv \frac{\partial \ln Y_t}{\partial \ln K_t} + \frac{\partial \ln Y_t}{\partial \ln L_t} = 1 + \epsilon.$$

2.2.2 Internal Returns to Scale

Changes in firm-level output occur through three channels: changes in technology, changes in firm capital and changes in firm-level total labour. The total derivative of firm-level output is:

$$d \ln y_t(\jmath) = \frac{\partial \ln y_t(\jmath)}{\partial \ln A_t} d \ln A_t + \frac{\partial \ln y_t(\jmath)}{\partial \ln k_t(\jmath)} d \ln k_t(\jmath) + \frac{\partial \ln y_t(\jmath)}{\partial \ln \ell_t^{tot}(\jmath)} d \ln \ell_t^{tot}(\jmath) \tag{40}$$

Returns to scale at the firm-level is the response of firm output to a change in all inputs:

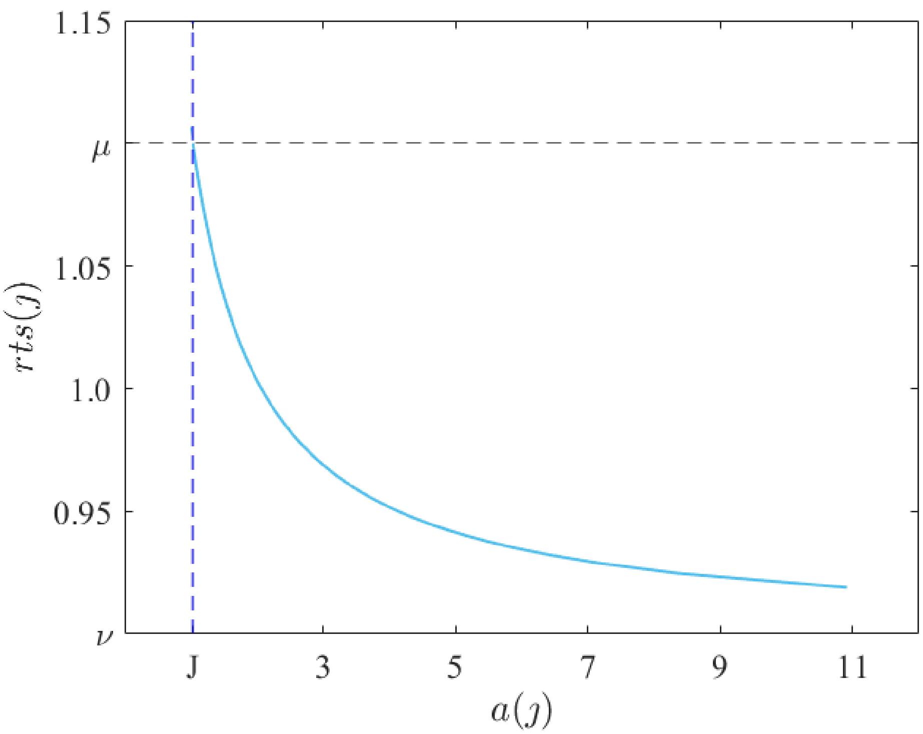
$$\text{RTS} \equiv \frac{\partial \ln y_t(\jmath)}{\partial \ln k_t(\jmath)} + \frac{\partial \ln y_t(\jmath)}{\partial \ln \ell_t^{tot}(\jmath)} = \nu \left(1 + (1-\alpha) \frac{\phi}{\ell_t(\jmath)} \right). \tag{41}$$

This follows from equations (9) and (10). This expression for returns to scale is endogenous. It depends on the non-production to production labour ratio and this will differ by firm according to their productivity draw. We can rearrange (41) to show:

$$RTS_t(\jmath) = \nu + (\mu - \nu) \frac{a(J_t)}{a(\jmath)}. \tag{42}$$

Returns to scale are bounded between variable returns to scale and the markup $RTS_t(\jmath) \in (\nu, \mu)$. Figure 1 plots firm-level returns to scale as a function of $a(\jmath)$ as given in equation (42). More productive firms have lower returns to scale. The cut-off firm $a(\jmath) = a(J_t)$ has the highest level of returns to scale which is equal to the markup, and returns to scale converge on variable returns to scale ν for large firms as the fixed cost becomes negligible.

Figure 1: Firm-level Returns to Scale



Returns to scale in variable production measures the response of firm output to variable inputs:

$$\text{Variable RTS} \equiv \frac{\partial \ln y_t(\jmath)}{\partial \ln k_t(\jmath)} + \frac{\partial \ln y_t(\jmath)}{\partial \ln \ell_t(\jmath)} = \nu. \tag{43}$$

We refer to this as variable returns to scale.

2.2.3 Average RTS

If we consider returns to scale of the firm with average productivity $\bar{a}(\bar{\jmath})$, under Pareto productivity distribution, then

$$RTS(\bar{\jmath}) = \mu - \frac{1-\nu}{\vartheta}, \tag{44}$$

which is increasing in μ, ν, decreasing in ϑ. Returns to scale of the average firm tends towards the markup (which is the returns to scale of the smallest, cut-off, firm) as ϑ gets large. This is because a higher Pareto shape parameter implies a higher density in the left-hand side of the distribution and a thinner right-hand tail. This implies a higher share of low-productivity firms which have high RTS, close to the markup. Average returns to scale of operating firms is independent of the overhead fixed cost ϕ. Although changes in the fixed cost ϕ affect firm-level returns to scale, $RTS(\jmath)$, on average returns to scale do not change as J_t adjusts. For example, a higher fixed cost raises selection J_t such that operating firms have a higher level of production labour and the overhead cost to production labour ratio in (41) remains unchanged. In other words, the overhead labour to production labour ratio is independent of the fixed cost:

$$\frac{\phi}{\ell(\bar{\jmath})} = \frac{\vartheta(\mu - \nu) - (1 - \nu)}{\nu(1 - \alpha)\vartheta}.$$

Although invariant to fixed costs, average returns to scale are increasing in variable returns to scale ν:

$$\frac{dRTS(\bar{\jmath})}{d\nu} = \frac{1}{\vartheta} > 0.$$

2.3 Reducing the model

We can reduce the model to two dynamic equations in two unknown variables $\{C_t, K_t\}$. With a Pareto distribution on $A(\jmath)$, labour utilized for production u is a function of exogenous parameters, consequently we drop the time subscript on utilization in the model under Pareto:

$$u = \left[1 + \frac{\mu - \nu}{\nu(1 - \alpha)}\left(1 - \frac{1}{\vartheta}\left(\frac{1 - \nu}{\mu - \nu}\right)\right)\right]^{-1}. \tag{45}$$

With utilization in terms of exogenous parameters, then TFP is a function of exogenous parameters and J_t:

$$\begin{aligned} TFP_t(J_t) &= \left(\frac{1 - u}{\phi}\right)^{1+\epsilon-\nu} u^{(1-\alpha)\nu} \bar{a}(J_t)^{\mu-\nu} & (46)\\ &= \left(\frac{1 - u}{\phi}\right)^{1+\epsilon-\nu} u^{(1-\alpha)\nu} \left(\frac{\vartheta}{\vartheta - \Gamma}\right)^{\mu-\nu} (1 - J_t)^{-\frac{1-\nu}{\vartheta}}. & (47)\end{aligned}$$

In turn aggregate output is a function of (J_t, K_t):

$$Y_t(J_t, K_t) = TFP_t(J_t) K_t^{\alpha\nu}. \tag{48}$$

If we combine labour demand which determines wage with free entry which equates entry cost to expected profit, and then use the expression for aggregate output $Y_t(J_t, K_t)$, then we can derive a relationship between the productivity cut-off J_t and capital K_t:

$$1 - J_t = \left[\frac{\kappa \mu u^{1-(1-\alpha)\nu} (\vartheta - \Gamma)^{1-\nu+\mu}}{\phi^{\nu-\epsilon} \nu (1-\alpha) \Gamma (1-u)^{1+\epsilon-\nu} \, \vartheta^{\mu-\nu} K_t^{\alpha\nu}} \right]^{\frac{\vartheta}{\vartheta-(1-\nu)}} \tag{49}$$

Using this expression, TFP can be expressed in terms of capital and in turn aggregate output can be expressed as $Y_t(K_t)$. Using these expressions gives the rental rate as a function of capital:

$$r_t = \alpha \frac{\nu}{\mu} \frac{Y_t(K_t)}{K_t}. \tag{50}$$

Therefore the system reduces to a two-dimensional dynamical system:

$$Y_t(K_t) - C_t = K_{t+1} - (1-\delta)K_t \tag{51}$$

$$\left(\frac{C_{t+1}}{C_t} \right)^{\sigma} = \beta(r_{t+1}(K_{t+1}) + (1-\delta)). \tag{52}$$

2.4 Steady state

A steady-state equilibrium is an equilibrium in which prices, quantities, entry and firm productivity are constant. In steady state, the consumption Euler equation implies that

$$\tilde{r} = \frac{1}{\beta} + (1-\delta). \tag{53}$$

In turn the rental rate equation yields a steady state expression for capital $\tilde{K}$ which leads to a steady state level of consumption from the capital accumulation equation. With $\{\tilde{C}, \tilde{K}\}$ all other endogenous model variables are determined in steady state. We can show that the cut-off level of productivity is:

$$1-\tilde{J} = \left[\mu\left(\frac{\kappa}{1-\nu}\right)^{1-\alpha\nu}\left(\frac{\tilde{r}}{\alpha\nu}\right)^{\alpha\nu}\left[\frac{\mu-\nu}{\nu(1-\alpha)\vartheta}\right]^{(1-\alpha)\nu}(\vartheta-\Gamma)^{\mu-\alpha\nu}\,\phi^{\epsilon-(1-\alpha)\nu}\,(1-u)^{-\epsilon}\,\vartheta^{1-\mu+(1-\alpha)\nu}\right]^{\frac{\vartheta}{\vartheta(1-\alpha\nu)-(1-\nu)}} \tag{54}$$

This is a function of $\mu, \nu, \alpha, \beta, \delta, \kappa, \phi, \vartheta$, and $\Gamma, u, \tilde{r}$ are defined above.

2.5 Comparative Statics

How does average productivity respond to changes in exogenous technical parameters? Changes in exogenous parameters such as ν and ϕ, denoted x below, affect average productivity through the degree of selection:

$$\frac{d\ln\bar{A}(\tilde{J})}{d\ln x} = \frac{\partial\ln\bar{A}}{\partial\ln(1-\tilde{J})} \times \frac{d\ln(1-\tilde{J})}{d\ln x} = -\frac{1}{\vartheta} \times \frac{d\ln(1-\tilde{J})}{d\ln x}. \tag{55}$$

The first term captures that greater selection ($1-\tilde{J} \to 0$) always increases average productivity. This follows from equation (38). The second term captures how selection (54) responds to a change in an exogenous parameter.[3]

How does selection respond to a change in variable returns to scale? The steady-state expression of (inverse) selection $1-\tilde{J}$ is nonlinear in ν. We cannot provide a closed-form result. Figure 2 shows cut-off $\tilde{J}$ for values of ν from 0.75 to 0.90. Exogenous parameters are set as in Table 1.

[3] In the appendix we present an analysis of ϕ. It has an ambiguous effect on selection and therefore productivity. The ambiguity depends on the degree of external returns to scale ϵ. For common restrictions (CES and no external returns), fixed costs increase selection and raise average productivity.

Table 1: Parameter Values for Comparative Statics

	Parameter	Value	Target
μ	Markup estimate	1.1	Hwang, Savagar, and Kariel (2022)
α	Capital share	0.25	
ϵ	External RTS	0.05	Gouel and Jean (2021)
κ	Entry cost	0.1	Barseghyan and DiCecio (2011)
β	Discount rate	0.96	Real interest rate
δ	Depreciation rate	0.08	Office for National Statistics
ϑ	Pareto shape	1.3	Match firm distribution
ϕ	Overhead cost	0.5	Match share active firms

The markup estimate matches evidence from the UK, as in Hwang, Savagar, and Kariel (2022). The value of α implies a capital share of 0.25ν. The value of κ follows Barseghyan and DiCecio (2011). Love of variety ϵ is set to a value from Gouel and Jean (2021). The depreciation rate δ is based on a weighted average from ONS data. The discount factor β is chosen to match the average real interest rate of 2.08 over the period, from the equation $\tilde{r} = \frac{1}{\beta} + 1 - \delta$.[4] The Pareto shape ϑ is calibrated to match the firm size distribution, as in Table 13 in the Appendix. The overhead cost ϕ is set such that the proportion of active and inactive firms $(\tilde{J})$ is empirically plausible.

[4] Data on UK long-term government bond and inflation used to compute the real interest rate from FRED database: https://fred.stlouisfed.org/series/IRLTLT01GBM156N and https://fred.stlouisfed.org/series/FPCPITOTLZGGBR.

Figure 2: Cut-off Productivity $\tilde{J}$ and variable returns to scale ν.

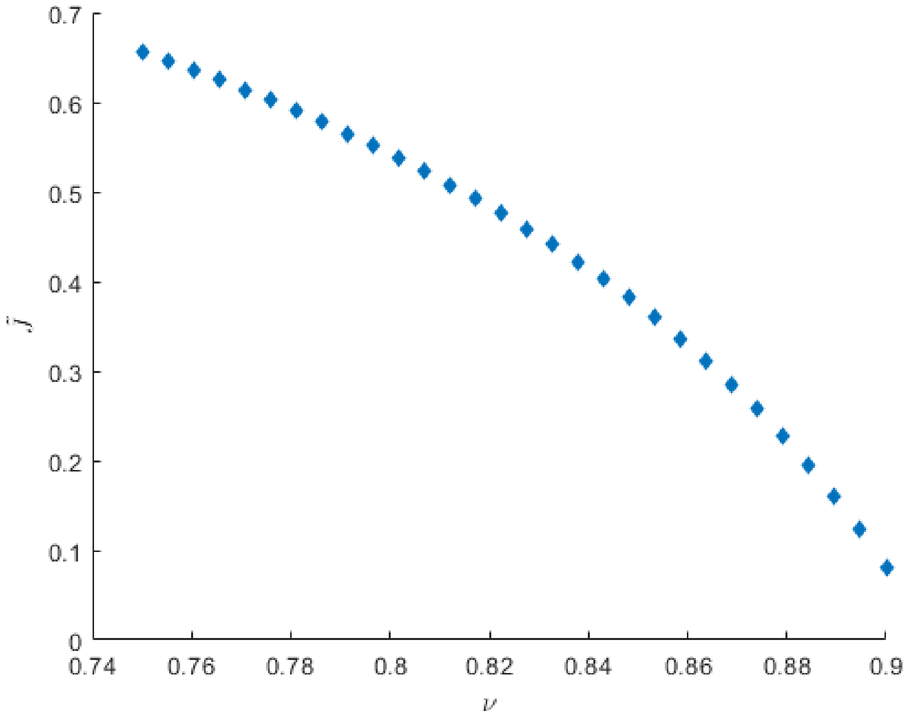
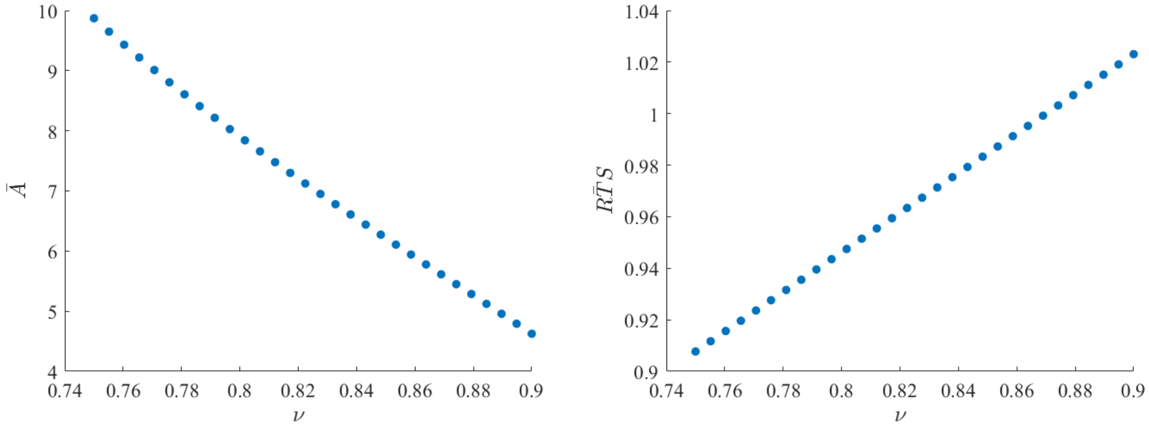


Figure 2 shows that the cut-off productivity is declining in variable returns to scale ν. This occurs because the wage declining in ν, such that payments to overheads are falling in ν, therefore less productive firms can pay the fixed cost and survive with higher ν. The negative relationship between the equilibrium wage and ν occurs because it reduces firm entry which in turn lowers the wage. A higher ν reduces expected profits via Equation (123), which decreases entry and hence the number of operating firms.[5] This lowers aggregate output, which puts downwards pressure on the equilibrium wage.

Figure 3 plots average productivity and average returns to scale for ν between 0.75 and 0.90. It is clear that a rise in ν lowers average productivity, but raises average returns to scale.

[5] The number of operating firms $N = \frac{1-u}{\phi}$. A rise in ν will lower $1-u$, by Equation (45). Hence the number of operating firms falls in ν.



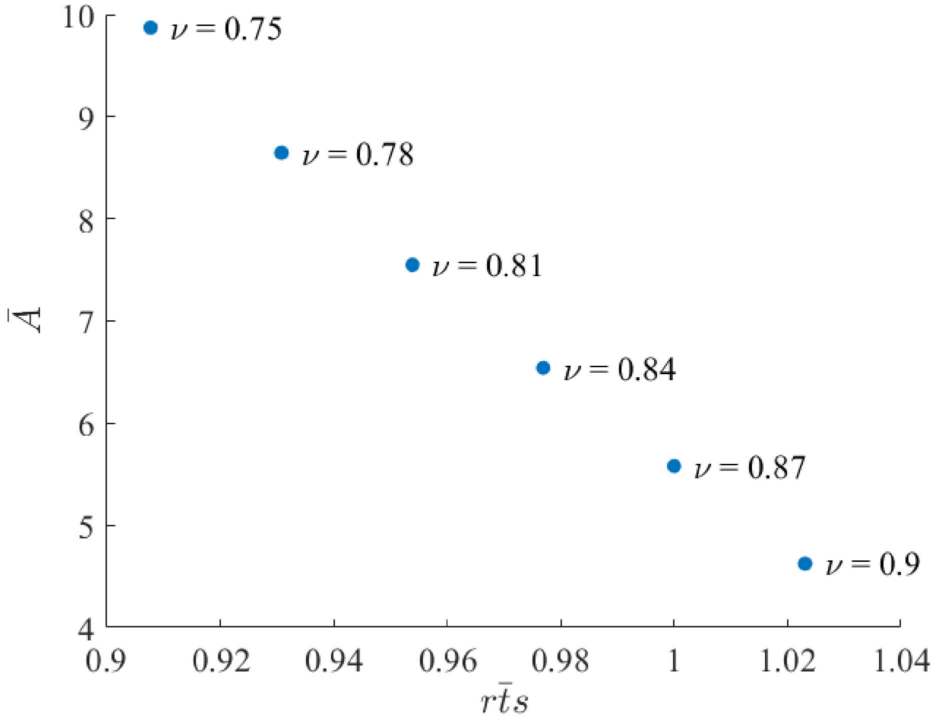
(a) Industry equilibrium average productivity for values of variable returns to scale ν.

(b) Industry equilibrium average returns to scale for values of variable returns to scale ν.

Figure 3: Equilibrium effect of a change in ν on endogenous averages of productivity and returns to scale.

The model produces the following result. As ν rises, the returns to scale of the average firm increases, given Equation (44). However there is *less* selection with higher ν, so average productivity will fall. This gives a negative relationship between returns to scale and productivity across industries characterised by variation in ν. This is plotted in Figure 4.

Figure 4: Scatter plot of average productivity and average returns to scale.



3 Data & Estimation

We use data from the UK's annual production survey the ARDx 1998-2014. This is a firm-level panel dataset that covers all large firms and a representative sample of small firms stratified by size, sector, region. Large firms are surveyed annually, whereas small firms are surveyed for a fixed number of years. The data contains 60,000 firms each year, 11 million workers and two-thirds of gross value added. Firms report a range of production data, our main interest is output, labour, materials, and investment which we use to construct capital. For output, we observe both value-added and gross output. We focus on value-added in our primary analysis, but use gross output for robustness checks with other production function estimation strategies.

In the appendix we provide details of our data cleaning, deflation, capital construction, SIC code matching and summary statistics for regression data.

3.1 Production Function Estimation

To stay close to our theory section, we estimate Cobb-Douglas production functions on value-added output. Our main analysis uses the Ackerberg, Caves, and Frazer (2015) (ACF) estimation methodology, and for robustness we provide results for Olley and Pakes (1996) (OP), Levinsohn and Petrin (2003) (LP) and Gandhi, Navarro, and Rivers (2020) (GNR). All these estimation methodologies address the perennial problem in production function estimation that there is omitted variable bias: input variables are correlated with unobserved firm-level productivity. With Cobb-Douglas production functions we obtain a single, time-invariant, coefficient for each input in the production function. This coefficient represents the elasticity of firm output to an input for the average firm in the panel over the period we estimate. We provide additional detail on ACF and GNR approaches in Appendix B. [6]

[6] In Kariel, Mainente, and Savagar (2022) we provide a comprehensive, reduced-form, empirical study of returns to scale in the UK economy. We consider the various choices involved in estimating production functions, such as functional form (translog or Cobb-Douglas), output measure (gross output or value-added) and estimation technique. This is an extensive exercise beyond the scope of this paper. In this paper our main goal is to be as consistent with our theoretical model.

3.2 Mapping Model Production Function to the Data

In the data we observe the following variables at the firm-level: revenue $p_t(\jmath)y_t(\jmath)$, capital $k_t(\jmath)$ and total labour $\ell_t^{\text{tot}}(\jmath)$. We cannot distinguish between production labour $\ell_t(\jmath)$ and overhead labour ϕ. Hence, we cannot estimate the firm-level production function (9) directly. If we were able to do this, then the sum of the coefficients would give us the response of revenue to variable inputs. Instead, we estimate the response of revenue to a change in all inputs. That is, we estimate the following regression:

$$\ln p_t(\jmath)y_t(\jmath) = \beta_1 \ln k_t(\jmath) + \beta_2 \ln \ell_t^{\text{tot}}(\jmath) + \varepsilon_t(\jmath). \tag{56}$$

3.2.1 Estimated Revenue Elasticity

The coefficients β_1 and β_2 represent the elasticity of firm-level revenue to firm-level capital and firm-level total labour, respectively. The coefficients are treated as common across all $\jmath = 1 \dots N$ within each industry and common across time. The coefficients are

$$\beta_1 = \frac{\partial \ln p_t(\jmath)y_t(\jmath)}{\partial \ln k_t(\jmath)} = \frac{\nu}{\mu}\alpha \tag{57}$$

$$\beta_2 = \frac{\partial \ln p_t(\jmath)y_t(\jmath)}{\partial \ln \ell_t^{\text{tot}}(\jmath)} = \frac{\nu}{\mu}(1-\alpha)\left(1 + \frac{\phi}{\ell(\bar{\jmath})}\right) = \frac{1}{\mu}\left(\mu - \nu\alpha - \frac{1-\nu}{\vartheta}\right). \tag{58}$$

Where $\ell(\bar{\jmath})$ reflects that pooling firms within an industry and time will yield average firm production labour. If we had true output measures rather than revenue then there would be no $1/\mu$ on the right-hand side of both equations. If we had a measure of production labour then there would be no $\left(1 + \frac{\phi}{\ell_t(\jmath)}\right)$ in Equation (58). Therefore the sum of the regression coefficients give revenue elasticity which is the change in output from a change in all inputs (output elasticity) divided by the markup:

$$\beta_1 + \beta_2 = \frac{\nu}{\mu}\left(1 + (1-\alpha)\frac{\phi}{\ell(\bar{\jmath})}\right) = \frac{1}{\mu}\left(\mu - \frac{1-\nu}{\vartheta}\right) = 1 - \frac{1-\nu}{\mu\vartheta}. \tag{59}$$

3.2.2 Estimated Technical Efficiency

We label the *revenue function residual* $\varepsilon_t(\jmath)$ for firm $\jmath$ in year t as revenue total factor productivity $\ln TFPR_t(\jmath)$:

$$\ln T\hat{FPR}_t(\jmath) = \ln p_t(\jmath) y_t(\jmath) - \hat{\beta}_1 \ln k_t(\jmath) - \hat{\beta}_2 \ln \ell_t^{\text{tot}}(\jmath). \tag{60}$$

Hat notation represents our estimated values. Specifically, $\hat{\beta}_1, \hat{\beta}_2$ are the estimated coefficients (revenue elasticities) for an average firm in the industry we run the regression on and $\ln T\hat{FPR}_t(\jmath)$ is the estimated revenue function residual for a specific firm given their capital, labour and revenue, teamed with industry average revenue elasticity coefficients. Our $\ln T\hat{FPR}_t(\jmath)$ measure is a good proxy of firm technical efficiency $A(\jmath)$, if overhead costs are small and if we purge the residual of industry-specific fixed effects which represent demand shocks. We provide a detailed discussion in the Appendix. Our main analysis focuses on variation across industries in average $\ln T\bar{FPR}(\imath)$ and $\hat{\beta}_1(\imath)+\hat{\beta}_2(\imath)$ where $\imath$ indexes an industry. The arithmetic average across all the firm-year residuals in an industry is

$$\ln T\bar{FPR} = \frac{\nu}{\mu}(1-\alpha)\left[\ln \ell_t(\bar{\jmath}) - \frac{\ell_t^{\text{tot}}(\bar{\jmath})}{\ell_t(\bar{\jmath})} \ln \ell_t^{\text{tot}}(\bar{\jmath})\right] + \frac{1-\nu}{\mu} \ln A(\bar{\jmath}) + \frac{\mu-1}{\mu} \ln Y_t + \frac{1+\epsilon-\mu}{\mu} \ln N_t. \tag{61}$$

By studying differences in $T\bar{FPR}$ across industries as a proxy for differences in $\bar{A}$ (which is what we study in our model) across industries, we are implicitly assuming the average labour component and two demand shocks are constant across industries.[7] In Appendix E, we show that $\ln T\bar{FPR}$ in our model is declining in ν, and that this is mostly driven by falling true average productivity $\bar{A}$. Therefore, $\ln T\bar{FPR}$ is a good proxy for $\ln \bar{A}$.

[7] We also run a regression specification to control for industry-specific shocks, and the main finding is unchanged.

3.2.3 Residual RTS correlation

We then seek to understand the relationship between the industry average $\ln T\bar{F}PR$ and the sum of the estimated coefficients, which represent returns to scale. To do this we run the regression:

$$\ln T\bar{F}PR(\imath) = \gamma_0 + \gamma_1 \ln\left[\hat{\beta}_1(\imath) + \hat{\beta}_2(\imath)\right] = \gamma_0 + \gamma_1 \ln\left[1 - \frac{1-\nu(\imath)}{\mu\vartheta}\right]. \tag{62}$$

for $\imath = 1 \ldots M$ where there are M 2-digit industries. We estimate the γ_1 coefficient to be negative which implies that an industry with higher $\hat{\beta}_1 + \hat{\beta}_2$ has lower TFPR on average. We have shown in our theory that only an increase in ν can both raise $\hat{\beta}_1 + \hat{\beta}_2$ and also decrease TFPR. Therefore the negative $\hat{\gamma}_1$ identifies ν as the parameter that changes across industries to drive the negative relationship between returns to scale and productivity.

4 Empirical Results

In this section, we present results on revenue elasticity in the UK economy between 1998 and 2014. These estimates are computed by summing the estimated coefficients from the production function estimation. The estimated coefficients are either labour, capital, and materials (gross output production functions) or labour and capital (value added production functions). We will refer to our estimates as ‘returns to scale’, although technically they are revenue elasticities. Given our identification strategy assumes CES demand and monopolistic competition, markups will be constant across firms and over time, so revenue elasticity will be proportional to returns to scale. Our main findings are:

1. On average, returns to scale in the UK is slightly above constant.

2. Returns to scale is heterogeneous across sectors (0.53 - 1.52).

3. Returns to scale in the UK has risen over time.

4. Returns to scale and TFPR are negatively related across industries.

Firstly, we present the estimates of returns to scale over the whole period.

Table 2: Returns to Scale: Cobb-Douglas production function, 1998 - 2014

	Olley and Pakes (1996)	Levinsohn and Petrin (2003)	Ackerberg, Caves, and Frazer (2015)	Gandhi, Navarro, and Rivers (2020)
		Economy-wide		
RTS	1.018	1.137	1.051	1.024
N	303,069	449,484	527,813	527,813
		Manufacturing		
RTS	1.252	1.121	1.143	1.034
N	95,424	123,552	120,712	120,712
		Construction		
RTS	1.025	0.805	1.192	1.044
N	22,123	50,172	51,784	51,784
		Wholesale/Trade/Transport		
RTS	1.027	1.009	0.926	1.016
N	74,988	129,043	181,985	181,985
		Services		
RTS	1.021	0.938	1.067	1.015
N	77,209	146,717	173,332	173,332

Table 2 presents estimates of average returns to scale for the whole economy and macro sectors, across different estimation methods. The underlying coefficients are contained in Table 9 in the Appendix. Estimates of average returns to scale in the U.K. from 1998 - 2014 are reasonably close in magnitude given the methodological differences and underlying assumptions on firm behaviour. Each estimate suggests returns to scale exceed one. This suggests that, on average, firms operate to the left of the *minimum efficient scale*, as average costs exceed marginal costs.

We believe that the ACF and GNR results are most reliable methods, as the former deals with the identification issue on labour, and the latter sidesteps this issue with use of the first-order condition on materials. These estimates suggest that average returns to scale in the UK from 1998 - 2014 is greater than unity, in the range of 1.02 - 1.05.

Table 2 also presents estimates of returns to scale at the sectoral level. We find that returns to scale is greatest in the UK in Manufacturing, and lowest in Services. The estimates following ACF and GNR show a clear split between returns to scale in Manufacturing and Construction compared to Wholesale/Trade/Transport and Services: the former sectors have higher scale than the latter. This is less clear with OP and LP estimates, although these methods indicate that Manufacturing has greater returns to scale than Services.

We also estimate returns to scale at the 2-digit industry level. These results are contained in Table 10 in the Appendix, and show a wide range of scale economies, from 0.54 to 1.52. The industry with the lowest returns to scale is Sewerage Services, while that with the highest is Furniture Manufacturing.

4.1 Rises in Returns to Scale

We estimate production functions on four shorter sub-periods, in order to track changes in returns to scale over time. Table 3 presents these estimates of returns to scale following ACF. Underlying coefficients are found in Table 11. Estimates with GNR are provided in the Appendix Table 12.

On aggregate, there is some evidence of a rise in scale economies over time. Estimates in the late 1990s suggest returns to scale below unity, but by the 2010s we find returns to scale above one.

Returns to scale has also increased across all macroeconomic sectors, although the rise has not been consistent. Table 3 also presents returns to scale in each sub-period, for each macro sector. Average returns to scale is higher in each sector when estimated between 2010 - 2014, compared to 1998 - 2001. The greatest rise in returns to scale is found in Construction and Services, from 0.91 and 1.02 to 1.29 and 1.11 respectively.

Table 3: Changing Returns to Scale, 1998 - 2014.

	1998 - 2001	2002 - 2005	2006 - 2009	2010 - 2014
	Economy-Wide			
RTS	0.988	1.081	1.046	1.061
N	153,874	144,465	108,619	120,855
	Manufacturing			
RTS	1.11	1.24	1.10	1.15
N	41,572	36,074	24,280	21,626
	Construction			
RTS	0.91	0.87	1.08	1.29
N	13,050	13,180	9,797	14,145
	Wholesale/Trade/Transport			
RTS	1.13	1.04	1.00	1.17
N	32,792	31,360	27,476	37,415
	Services			
RTS	1.02	1.03	1.02	1.11
N	34,698	34,241	32,070	45,708

All estimates follow Ackerberg, Caves, and Frazer (2015) with a value-added Cobb-Douglas production function.

The rise in returns to scale over time is more apparent when we estimate at the 2-digit industry level. Figure 5 plots a comparison of returns to scale in 1998 - 2001, compared to 2010 - 2014, across sectors, using the ACF estimation.[8] GNR estimates are provided in Figure 14 in the Appendix. It is clear that most industries experienced an increase in returns to scale, as the majority of points sit above the 45 degree line.

[8] We remove industries where estimated factor elasticities are below zero or above one.

Figure 5: Changing Returns to Scale by 2-digit SIC, ACF Estimation.

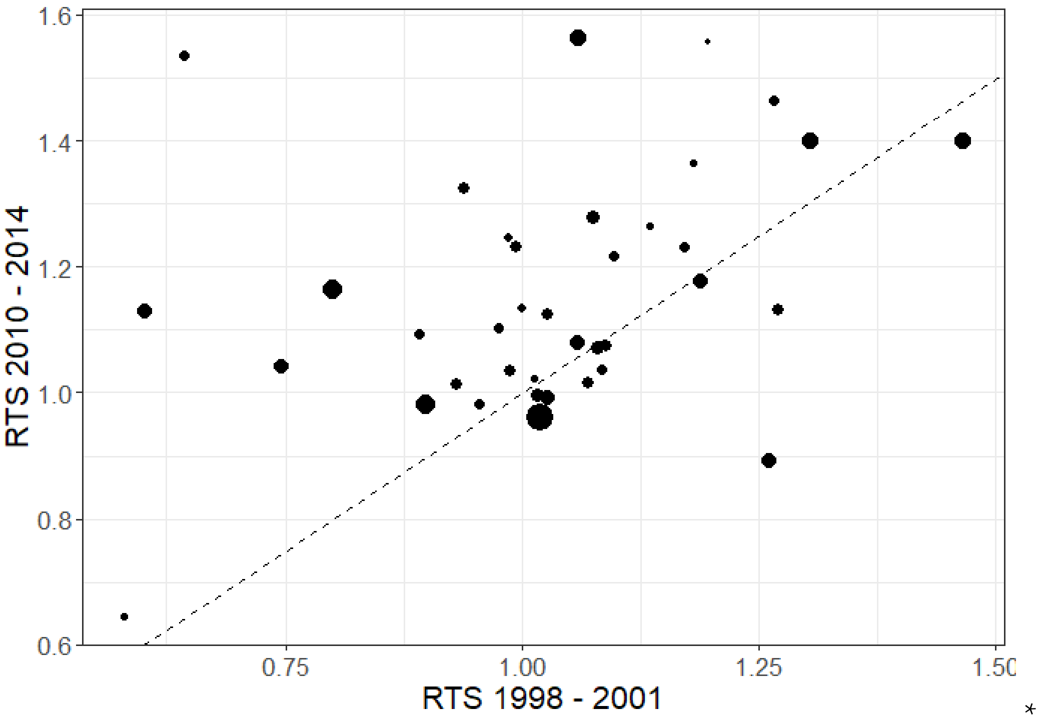


Figure 6: Comparison of returns to scale at 2-digit SIC level, from 1998 - 2001 to 2010 - 2014. Size of points represents the average number of firms in that sector in each period. Dotted line is 45 degree line: points above that line are consistent with a rise in returns to scale.

4.2 Returns to Scale & Productivity

We estimate TFPR using control function methods. Figure 7 presents average log TFPR across all firms in each year, when the production function is estimated following Ackerberg, Caves, and Frazer (2015). We see a rise in productivity until 2004, followed by a faster rise until 2009. It falls sharply before recovering somewhat over the following years. This result is robust across estimation methods (Gandhi, Navarro, and Rivers 2020), although the levels differ slightly (see Figure 12 in the Appendix).

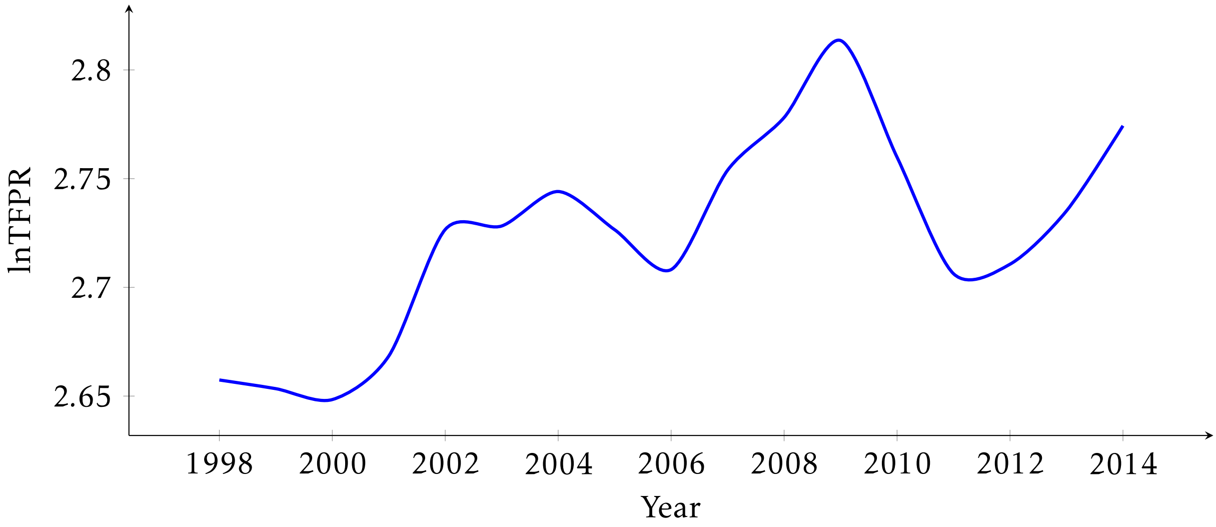


Figure 7: Aggregate TFPR (ACF)

At the aggregate level, both returns to scale and productivity have risen over time. We want to investigate the relationship across industries, as in our theoretical framework. Figure 8 contains a scatter plot of returns to scale and average log TFPR across 2-digit sectors for ACF estimation. We obtain the same general result with GNR estimation, plotted in Figure 16 in the Appendix.

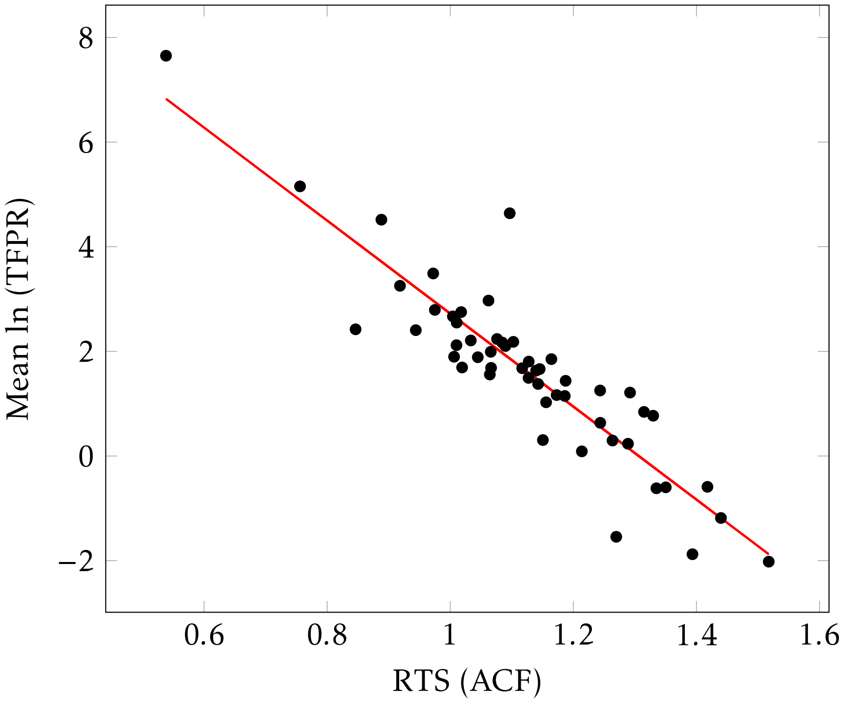


Figure 8: Relationship between Revenue Elasticity and TFP, ACF

This evidence shows that industries with higher returns to scale have lower productivity. This is consistent with our model, which characterises different industries by varying levels of ν.

Table 4 contains estimates from a regression of returns to scale on average log TFPR across 2-digit sectors. All estimates are from a production function following Ackerberg, Caves, and Frazer (2015). We separate the sample into sub-periods to understand how this relationship has changed over time. Our results suggest there is a strong negative relationship between returns to scale and productivity, and that this negative relationship has strengthened over time. An increase in average industry TFPR of 10% was associated with a fall in returns to scale of 0.014 in the late 1990s. By the early 2010s, such a rise in TFPR would suggest returns to scale is lower by 0.025.

Table 4: Regression: Returns to Scale and Productivity at Industry Level

	Dependent variable: Returns to Scale			
	1998 - 2001	2002 - 2005	2006 - 2009	2010 - 2014
Mean log TFPR	-0.143^{***}	-0.164^{***}	-0.234^{***}	-0.245^{***}
	(0.015)	(0.014)	(0.048)	(0.037)
N	60	60	62	62

*Note: Returns to scale estimated with a value-added Cobb-Douglas production function following Ackerberg, Caves, and Frazer (2015), estimated at the 2-digit SIC level. Estimates statistically significant at levels of 1%: ***, 5%: **, 10%: *. Number of sectors is determined by those for which a production function can be estimated in the sub-period, with estimated elasticities between zero and one.*

According to Section 3.2, our measure of TFPR contains both the average of true firm-level productivity and a sector-specific demand shock. In order to control for such sector-specific shocks, we run a regression of returns to scale on log TFPR with fixed-effects for the sector and time period. Effectively, we are combining the observations used for the regressions in each column of Table 4, and running a fixed-effects regression. The results are presented in Table 5. Column (1) presents an estimated coefficient from a simple pooled regression. The other three columns introduce weights (the number of firms used for estimation in each sector × period), 2-digit sector fixed-effects, and then period fixed-effects. The negative relationship between returns to scale and log productivity is negative and strongly statistically significant across all specifications. We believe that this provides evidence that sector-specific shocks are not driving our results.

Table 5: Regression: Returns to Scale and Productivity at Industry Level

	Dependent variable: Returns to Scale			
	(1)	(2)	(3)	(4)
Mean log TFPR	-0.135^{***}	-0.143^{***}	-0.150^{***}	-0.165^{***}
	(0.009)	(0.010)	(0.017)	(0.008)
Weighted (# firms)		✓	✓	✓
2-digit SIC FE			✓	✓
Period FE				✓
N	214	214	214	214
R^2	0.614	0.647	0.808	0.926

*Note: Returns to scale estimated with a value-added Cobb-Douglas production function following Ackerberg, Caves, and Frazer (2015), estimated at the 2-digit SIC level. Estimates statistically significant at levels of 1%: ***, 5%: **, 10%: *. Weighted by the number of firms used to estimate returns to scale and TFPR in each sector × period. Periods are: 1998 - 2001, 2002 - 2005, 2006 - 2009, 2010 - 2014. Number of sectors is determined by those for which a production function can be estimated in the sub-period, with estimated elasticities between zero and one.*

In sum, we provide estimates of returns to scale in the UK and find that they are slightly above constant on average. There is significant sectoral heterogeneity, with higher returns to scale in Manufacturing than Services. We find that returns to scale has increased between 1998 and 2014, with the greatest rise occurring in the early 2000s, and driven by an increase in Construction and Services. This rise has occurred across the vast majority of 2-digit sectors. Finally, we show a strong negative relationship between returns to scale and TFPR, which has become stronger over time, and holds even when we control for sectors and time periods.

5 Conclusion

We show that considering both marginal and fixed costs drivers of returns to scale is important for productivity outcomes. Our theory implies that rising returns to scale in variable production can lead to a negative relationship between returns to scale and productivity, whereas changing fixed costs cannot cause this outcome. Using data for the UK from 1998 - 2014, we estimate firm-level returns to scale and productivity. We document a negative relationship between the two variables which is consistent with our theory. Overall, our results suggest a channel to explain the puzzling observation that in many advanced economies returns to scale are rising whilst productivity is stagnating. This is difficult to understand in many models as typically greater returns to scale increases selection of high productivity firms leading to rising returns to scale and productivity. We stress that the source of returns to scale matters to overturn this result.

References

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer (2015). "Identification Properties of Recent Production Function Estimators". In: *Econometrica* 83.6, pp. 2411–2451.

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li (Nov. 2019). *A Theory of Falling Growth and Rising Rents*. NBER Working Papers 26448. National Bureau of Economic Research, Inc.

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt (2005). "Competition and innovation: An inverted-U relationship". In: *The Quarterly Journal of Economics* 120.2, pp. 701–728.

Asturias, Jose, Sewon Hur, Timothy J. Kehoe, and Kim J. Ruhl (2022). "Firm Entry and Exit and Aggregate Growth". In: *American Economic Journal: Macroeconomics*.

Atkeson, Andrew and Patrick J Kehoe (2005). "Modeling and measuring organization capital". In: *Journal of political Economy* 113.5, pp. 1026–1053.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen (2017). "Concentrating on the Fall of the Labor Share". In: *American Economic Review* 107.5, pp. 180–85.

Axtell, Robert L (2001). "Zipf distribution of US firm sizes". In: *science* 293.5536, pp. 1818–1820.

Baqaee, David, Emmanuel Farhi, and Kunal Sangani (Aug. 2021). *The Darwinian Returns to Scale*. Working Paper 27139. National Bureau of Economic Research.

Barkai, Simcha (2020). "Declining labor and capital shares". In: *The Journal of Finance* 75.5, pp. 2421–2463.

Barnett, Alina, Sandra Batten, Adrian Chiu, Jeremy Franklin, and Maria Sebastia-Barriel (2014). "The UK productivity puzzle". In: *Bank of England Quarterly Bulletin*, Q2.

Barseghyan, Levon and Riccardo DiCecio (Oct. 2011). "Entry costs, industry structure, and cross-country income and TFP differences". In: *Journal of Economic Theory* 146.5, pp. 1828–1851.

Barseghyan, Levon and Riccardo DiCecio (2016). “Externalities, endogenous productivity, and poverty traps”. In: *European Economic Review* 85.C, pp. 112–126.

Bessen, James (2020). “Industry Concentration and Information Technology”. In: *The Journal of Law and Economics* 63.3, pp. 531–555.

Collard, Fabrice and Omar Licandro (July 2021). “The neoclassical model and the welfare costs of selection”. In: *Working Paper (Jul 2021)*.

De Loecker, Jan, Jan Eeckhout, and Simon Mongey (2021). *Quantifying market power and business dynamism in the macroeconomy*. Tech. rep. Working paper (Sept 2021).

De Ridder, Maarten (Mar. 2019). *Market Power and Innovation in the Intangible Economy*. Discussion Papers 1907. Centre for Macroeconomics (CFM).

Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda (Dec. 2020). “Changing Business Dynamism and Productivity: Shocks versus Responsiveness”. In: *American Economic Review* 110.12, pp. 3952–90.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (May 2021). *How Costly Are Markups?* NBER Working Papers 24800. National Bureau of Economic Research, Inc.

Eggertsson, Gauti B., Jacob A. Robbins, and Ella Getz Wold (2021). “Kaldor and Piketty’s facts: The rise of monopoly power in the United States”. In: *Journal of Monetary Economics* 124. The Real Interest Rate and the MarginalProduct of Capital in the XXIst CenturyOctober 15-16, 2020, S19–S38.

Foster, Lucia S, John C Haltiwanger, and Cody Tuttle (Sept. 2022). *Rising Markups or Changing Technology?* Working Paper 30491. National Bureau of Economic Research.

Gandhi, Amit, Salvador Navarro, and David A. Rivers (2020). “On the Identification of Gross Output Production Functions”. In: *Journal of Political Economy* 128.8, pp. 2973–3016.

Gao, Wei and Matthias Kehrig (July 2021). “Returns to Scale, Productivity and Competition: Empirical Evidence from US Manufacturing and Construction Establishments”. In: *Working Paper (Jul 2021)*.

Girma, S. and H. Görg (2002). "Foreign Ownership, Returns to Scale and Productivity: Evidence from UK Manufacturing Establishments". In: *CEPR Discussion Paper Series*.

Gouel, Christophe and Sebastien Jean (2021). *Love of Variety and Gains from Trade*. Working Papers. CEPII research center.

Gutierrez, German, Callum Jones, and Thomas Philippon (Feb. 2019). *Entry Costs and the Macroeconomy*. Working Paper 25609. National Bureau of Economic Research.

Gutiérrez, Germán, Callum Jones, and Thomas Philippon (2021). "Entry costs and aggregate dynamics". In: *Journal of Monetary Economics* 124. The Real Interest Rate and the MarginalProduct of Capital in the XXIst CenturyOctober 15-16, 2020, S77–S91.

Harris, Richard and Eunice Lau (Apr. 1998). "Verdoorn's law and increasing returns to scale in the UK regions, 1968–91: some new estimates based on the cointegration approach". In: *Oxford Economic Papers* 50.2, pp. 201–219.

Hopenhayn, Hugo (2014). "Firms, misallocation, and aggregate productivity: A review". In: *Annual Review of Economics* 6.1, pp. 735–770.

Hopenhayn, Hugo and Richard Rogerson (1993). "Job Turnover and Policy Evaluation: A General Equilibrium Analysis". In: *Journal of Political Economy* 101.5, pp. 915–938.

Hwang, Kyung In, Anthony Savagar, and Joel Kariel (2022). *Market Power in the UK*. Working Papers.

Kariel, Joel, Joao Mainente, and Anthony Savagar (2022). *Returns to Scale in the UK*. Working Papers.

Kim, Daisoon (2021). "Economies of scale and international business cycles". In: *Journal of International Economics* 131, p. 103459.

Lashkari, Danial, Arthur Bauer, and Jocelyn Boussard (2019). *Information Technology and Returns to Scale*. 2019 Meeting Papers 1380. Society for Economic Dynamics.

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

Licandro, Omar (2022). “Innovation and Growth: Theory”. In: *Macroeconomic Modelling of R&D and Innovation Policies*. Ed. by Ufuk Akcigit, Cristiana Benedetti Fasil, Giammario Impullitti, Omar Licandro, and Miguel Sanchez-Martinez. Cham: Springer International Publishing, pp. 23–61.

Luttmer, Erzo G. J. (Aug. 2007). “Selection, Growth, and the Size Distribution of Firms”. In: *The Quarterly Journal of Economics* 122.3, pp. 1103–1144.

Martin, Ralf (2002). *Building the capital stock*. CeRiBA Working Paper. The Centre for Research into Business Activity.

Olley, G. Steven and Ariel Pakes (1996). “The dynamics of productivity in the telecommunications equipment industry”. In: *Econometrica* 64.6, pp. 1263–1297.

Oulton, Nicholas (1996). “Increasing Returns and Externalities in UK Manufacturing: Myth or Reality?” In: *The Journal of Industrial Economics* 44.1, pp. 99–113.

Restuccia, Diego and Richard Rogerson (2008). “Policy distortions and aggregate productivity with heterogeneous establishments”. In: *Review of Economic dynamics* 11.4, pp. 707–720.

Riley, Rebecca, Chiara Rosazza-Bondibene, and Garry Young (June 2015). *The UK productivity puzzle 2008-13: evidence from British businesses*. Bank of England working papers 531. Bank of England.

Ruzic, Dimitrije and Sui-Jade Ho (Aug. 2019). “Returns to Scale, Productivity Measurement, and Trends in U.S. Manufacturing Misallocation”. In: *INSEAD working paper*.

Van Reenen, John (2018). “Increasing differences between firms: market power and the macro-economy”. In: *CEP Discussion Papers*.

Appendix

A Data

We use the Annual Respondents Database (ARD). Specifically, we use a version called the ARDx, where the 'x' suffix denotes that it is the time-series version of the ARD, rather than an independent annual release. The bulk of variables in the ARD come from the Annual Business Survey (ABS) which is an annual, mandatory, survey of firms in the UK economy. The ARD differs slightly from the ABS as the ONS add additional information from other surveys to the basic ABS data. Additional information is available at Annual Respondents Database, 1973-2008 and Annual Respondents Database X, 1998-2014. However, for us the most important documentation regards the ABS which is a core ONS product and is well-documented. The ARD is a research dataset bringing together the ABS and BRES, and prior to 2009 bringing together the two parts of the Annual Business Inquiry (ABI). The Annual Business Survey (ABS) Quality and Methodology Information report provides more information on limitations of the data and comparisons to related data. There is also an ABS Methodology web page which links to an ABS Technical Report.

A.1 Capital Construction

The Perpetual Inventory Method (PIM) allows construction of firm-level capital stocks when such data is unavailable, but investment data is present. The method here follows Martin (2002) and Hwang, Savagar, and Kariel (2022). The PIM is constructed using the following equation:

$$K_t = (1 - \delta)K_{t-1} + I_t.$$

K_t is the capital stock in period t, and I_t is investment in period t. However, to use this method, we need K_0 – the initial capital stock of a firm – which is not in this survey. To construct this series, each firm's K_0 is a revenue-weighted share of the industry-level capital stock in the first year that firm appears in the panel. Capital stock is then constructed for all future years with the above equation, with missing investment data interpolated. The depreciation rate is taken to be 18.195%, which is a weighted average of ONS depreciation rates for the three different capital categories: Building, Vehicles, Other.

A.1.1 Deflating

We convert firm gross output into real values using the ONS industry deflators.[9] Material inputs are deflated with the ONS producer price inflation data.[10] The capital stock is deflated with the ONS gross fixed capital formation deflator.[11]

A.1.2 Cleaning

For the purpose of our production function estimation, we exclude sectors: Agriculture, Public Sector, Finance & Insurance, Education, and Health.[12] We set out rules for SIC re-coding to ensure compatibility pre- and post-2007, when the classification is changed. For SIC codes post-2007, we divide the number by 1000 to match with pre-2007 codes. To reduce the influence of outliers, which may represent measurement or recording errors in the surveys, we winsorize firms with the top and bottom 0.1% of factor shares in revenue (M/Y, K/Y, L/Y) in each year. Table 6 contains number of firms at each stage of the data cleaning process, along with the final number of observations for estimation.

[9] https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/experimentalindustrydeflatorsuknonseasonallyadjusted

[10] https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/producerpriceindex

[11] https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/ybfu/ukea

[12] Standard Industrial Classification (SIC) 2007 codes: A, K, O, P, Q. These sectors were excluded from the survey after 2012. K,O,P were fully excluded and A,Q had various subsectors excluded.

Table 6: Data Cleaning: Firms Dropped

	# Firms
All ARD firms	854,732
Drop if no 2-digit sector	852,424
Drop if < 100 firms in sector	852,331
Drop non-market sectors	761,348
Take logs of regression variables	539,368
Drop outlier factor shares	527,813

A.2 Summary Statistics

Table 7 and 8 present descriptive statistics of the variables used in our regression analysis. It provides these for different broad industry groups, but we estimate the production function regressions at the 2-digit level.

Table 7: Descriptive Statistics of Regression Variables for Full Sample

	Mean	SD	p10	p50	p90	No. Obs
Revenue	39,736	675,183	92	1,458	42,797	527,813
Labour	224	2,213	2	20	349	527,813
Capital	7,696	150,007	22	351	7,915	527,813
Materials	29,651	636,176	32	703	26,255	527,813
Materials Share	0.55	-	0.17	0.58	0.87	527,813
Labour Share	0.26	-	0.04	0.23	0.52	527,813

Table 8: Descriptive Statistics of Regression Variables by Broad Sector

	Mean	SD	p10	p50	p90	No. Obs
Manufacturing						
Revenue	36,005	235,437	336	4,294	58,896	125,737
Labour	192	576	8	54	431	125,737
Capital	10,362	75,776	148	1,498	16,154	125,737
Materials	24,954	178,528	122	2,400	38,999	125,737
Materials Share	0.57	-	0.30	0.58	0.81	125,737
Labour Share	0.28	-	0.11	0.27	0.47	125,737
Construction						
Revenue	17,812	108,789	111	1,414	48,782	51,784
Labour	103	395	2	11	214	51,784
Capital	2,309	41,523	11	104	2,210	51,784
Materials	12,467	89,027	18	343	16,896	51,784
Materials Share	0.51	-	0.17	0.52	0.81	51,784
Labour Share	0.25	-	0.00	0.24	0.49	51,784
Trade, Wholesale, Transport						
Revenue	62,673	1,102,305	111	1,414	48,782	182,814
Labour	256	3,404	2	14	244	182,814
Capital	7,092	103,075	20	245	5,667	182,814
Materials	52,666	1,044,112	61	929	26,219	182,814
Materials Share	0.69	-	0.37	0.74	0.92	182,814
Labour Share	0.16	-	0.02	0.13	0.35	182,814
Services						
Revenue	25,276	284,335	65	728	28,673	179,028
Labour	249	1,627	2	17	403	179,028
Capital	8,821	228,905	20	218	5,435	179,028
Materials	14,417	209,297	15	242	11,263	179,028
Materials Share	0.41	-	0.09	0.38	0.77	179,028
Labour Share	0.34	-	0.06	0.32	0.68	179,028

B Production Function Estimation

OP, LP and ACF use the 'control function' approach. A key assumption of these methods is that the idiosyncratic productivity shock at time t does not affect the choice of state variables chosen by the firm in previous periods, but does affect the decision on free variables. GNR uses the firm's first-order condition to parametrically obtain the coefficient on the intermediate input of a gross output production function.

B.1 Control Function Approach (ACF)

For the control function approaches of OP, LP and ACF, we estimate a "value-added" or Leontief production function:

$$Y_{it} = \min\{Z_{it} K_{it}^{\beta_k} L_{it}^{\beta_l}, M_{it}^{\beta_m}\} e^{\epsilon_{it}}. \tag{63}$$

Taking logarithms yields the "value-added" production function:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \epsilon_{it} \tag{64}$$

where $\ln z_{it} = \beta_0 + \omega_{it}$. Firms draw productivity ω_{it} which is unobserved by the econometrician, leading to potential omitted variable bias, because the optimal firm input choices will be correlated with this variable.

We describe the control function method of Ackerberg, Caves, and Frazer (2015). The method makes assumptions on the timing of input choices to achieve identification, and uses materials expenditure as a proxy for unobserved productivity shocks. The following assumptions are required:

1. **Information Sets:** firms' information sets $\mathcal{I}_{it}$ include current and past productivity shocks $\{\omega_{i\tau}\}_{\tau=0}^{t}$, but firms know nothing about future shocks. The ex-post shocks η_{it} are expected to be zero on average: $\mathbb{E}\{\eta_{it}|\mathcal{I}_{it}\} = 0$.

2. **First-Order Markov Shocks:** productivity shocks follow a First-Order Markov Process, so $\omega_{it} = \mathbb{E}(\omega_{it}|\omega_{i,t-1}) + v_{it}$, and $\mathbb{E}\{v_{it}|\mathcal{I}_{it-1}\} = 0$.

3. **Timing of Input Choices:** firms accumulate capital according to $k_{it} = \kappa(k_{it-1}, i_{it-1})$ where investment i_{it-1} is chosen in period $t-1$. Labour l_{it} is chosen at period $t, t-1$ or in between. m_{it} is either chosen at the same time, or after l_{it} is chosen.

4. **Scalar Unobservable:** investment decisions $m_{it} = h_t(k_{it}, \omega_{it}, l_{it})$ have just one scalar unobservable ω_{it}, so there is no other across firm unobserved heterogeneity (e.g. adjustment costs, investment efficiency, input prices).

5. **Strict Monotonicity:** investment decisions are strictly monotonic in the scalar unobservable ω_{it}, so $m_{it} = h_t(k_{it}, \omega_{it}, l_{it})$.

Given that investment is strictly monotonic in the unobserved anticipated shock, this function can be inverted, and then substituted into the production function:

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + h_t^{-1}(k_{it}, m_{it}, l_{it}) + \eta_{it} = \Phi_t(k_{it}, \omega_{it}, l_{it}) + \eta_{it}.$$

This inverted function is unknown, so is approximated by a polynomial. Running the first-step regression yields an estimate of the composite term Φ_t. This exploits the moment condition $\mathbb{E}(\eta_{it}|\mathcal{I}_{it}) = 0$, because the ex-post productivity shock is unanticipated to the firm. Neither β_k nor β_l are identified in this stage, as both are contained in the composite term:

$$\widehat{\Phi}_t = \beta_k k_{it} + \beta_l l_{it} + \omega_{it}.$$

Production function parameters are estimated in the second stage. With the estimate of the composite term, estimates of the ex-ante productivity shock $\widehat{\omega}_{it}(\beta_k, \beta_l)$ can be computed for guesses of β_k, β_l. The implied $\widehat{\omega}_{it}(\beta_k, \beta_l)$ are non-parametrically regressed on their lag $\widehat{\omega}_{it-1}(\beta_k, \beta_l)$, and the residuals $\widehat{v}_{it}(\beta_k, \beta_l)$ are the implied innovations in productivity. The second stage moment condition is $\mathbb{E}(v_{it} + \eta_{it}|\mathcal{I}_{it-1}) = 0$. The sample analogue of the moment condition $\mathbb{E}\{v_{it}k_{it}\} = 0$ is:

$$\frac{1}{N}\frac{1}{T}\sum_i \sum_t \widehat{v}_{it}(\beta_k, \beta_l)k_{it} = 0.$$

If labour is assumed to be chosen after $t-1$, l_{it} will generally be correlated with v_{it}, so lagged labour is chosen as an additional moment condition. This procedure yields estimates $\widehat{\beta_k}, \widehat{\beta_l}$.

B.2 Cost Share Approach (GNR)

For GNR we estimate a gross output, Cobb-Douglas, production function of the following form:

$$Y_{it} = e^{\eta_{it}} e^{\omega_{it}} K_{it}^{\beta_k} L_{it}^{\beta_l} M_{it}^{\beta_m}$$

where $Y_{it}, K_{it}, L_{it}, M_{it}$ represent gross output, capital stock, employment, and materials respectively, while the β's are the production elasticities. ω_{it} are ex-ante shocks. η_{it} are ex-post shocks. Taking logarithms yields a regression equation with the log of gross output on the left-hand side, and logs of the factor inputs on the right-hand side:

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + \beta_m m_{it} + \omega_{it} + \eta_{it}.$$

Firms draw productivity ω_{it} which is unobserved by the econometrician. This causes omitted variable bias because optimal firm input choices depend on productivity.

In this section, we describe the Gandhi, Navarro, and Rivers (2020) cost share approach to estimating production functions. The production environment is the same as before, and the productivity process is also Markov. Capital and labour are chosen at $t-1$, while materials is chosen at t. Materials is strictly increasing in ex-ante productivity shock ω_{it}.

We take the first-order condition from choosing materials input to maximise profits to obtain the following relationship:

$$s_{it} = \ln\left(\frac{\partial f\left(l_{it}, k_{it}, m_{it}\right)}{\partial m_{it}}\right) + \ln \mathbb{E}(\eta_{it} | \mathcal{I}_{it}) - \eta_{it}$$

where $s_{it} = \ln \frac{P_t^M M_{it}}{P_t Y_{it}}$ is the log of the materials cost share in revenue. This can be written as:

$$s_{it} = \ln D(l_{it}, k_{it}, m_{it}) - \eta_{it} \tag{65}$$

Since $\mathbb{E}(\eta_{it}|\mathcal{I}_{it}) = \mathbb{E}(\eta_{it}|l_{it}, k_{it}, m_{it}) = 0$, Equation (65) is a regression equation of the log of the materials cost share in revenue on the inputs l_{it}, k_{it}, m_{it}.

Notice that the constant $\mathbb{E}(\eta_{it}|\mathcal{I}_{it})$ can be obtained because:

$$\mathbb{E}(\eta_{it}|\mathcal{I}_{it}) = \mathbb{E}\exp(\ln D(l_{it}, k_{it}, m_{it})) - s_{it}$$

Then the materials input elasticity can be identified:

$$\left(\frac{\partial f(l_{it}, k_{it}, m_{it})}{\partial m_{it}}\right) = \frac{D(l_{it}, k_{it}, m_{it})}{\mathbb{E}(\eta_{it}|\mathcal{I}_{it})} \tag{66}$$

Gandhi, Navarro, and Rivers (2020) propose regressing the log of the materials cost share on logs of a polynomial in capital, labour, and materials.

The rest of the production function can be identified as follows. Integrating over the materials input elasticity yields an expression which represents the part of the production function related to the materials input:

$$\int \frac{\partial f(l_{it}, k_{it}, m_{it})}{\partial m_{it}} dm_{it} = f(l_{it}, k_{it}, m_{it}) + g(k_{it}, l_{it}) \tag{67}$$

Combining this with the production function yields:

$$\mathcal{Y}_{it} = y_{it} - \eta_{it} - \int \frac{\partial f(l_{it}, k_{it}, m_{it})}{\partial m_{it}} dm_{it} = \omega_{it} - g(k_{it}, l_{it}) \tag{68}$$

where $\mathcal{Y}_{it}$ is a function of the data: output, minus the ex-post shock, minus the materials elasticity already obtained in the first stage. Gandhi, Navarro, and Rivers (2020) suggest using a polynomial sieve estimator so that the integral has a closed-form solution.

Using the Markov process on productivity, alongside the moment condition on the 'surprise' to ex-ante productivity $\mathbb{E}(v_{it}|k_{it}, l_{it}, \mathcal{Y}_{it-1}, k_{it-1}, l_{it-1}) = 0$, we obtain:

$$\mathbb{E}(\mathcal{Y}_{it}|k_{it}, l_{it}, \mathcal{Y}_{it-1}, k_{it-1}, l_{it-1}) = -g(k_{it}, l_{it}) + d(\mathcal{Y}_{it-1} + g(k_{it-1}, l_{it-1}) \tag{69}$$

so we regress $\mathcal{Y}_{it}$ on $k_{it}, l_{it}, \mathcal{Y}_{it-1}, k_{it-1}, l_{it-1}$.

The moment conditions for this stage are:

$$\mathbb{E}\left(\eta_{it}\widehat{\mathcal{Y}}_{it-1}\right) = 0$$

$$\mathbb{E}\left(\eta_{it}k_{it}^{\tau_k}l_{it}^{\tau_l}\right) = 0$$

where the second moment condition includes polynomials of capital and labour, up to τ_k, τ_l.

C Returns to Scale Estimates

Table 9: Elasticity Estimates: Cobb-Douglas production function, 1998 - 2014

	Olley and Pakes (1996)	Levinsohn and Petrin (2003)	Ackerberg, Caves, and Frazer (2015)	Gandhi, Navarro, and Rivers (2020)
		Economy-Wide		
β_l	0.497	0.635	0.545	0.329
β_k	0.521	0.501	0.505	0.181
β_m	-	-	-	0.514
N	303,069	449,484	527,813	527,813
		Manufacturing		
β_l	0.681	0.573	0.789	0.297
β_k	0.571	0.547	0.421	0.148
β_m	-	-	-	0.590
N	95,424	123,552	120,712	120,712
		Construction		
β_l	0.574	0.473	0.826	0.328
β_k	0.451	0.332	0.388	0.224
β_m	-	-	-	0.493
N	22,123	50,172	51,784	51,784
		Wholesale/Trade/Transport		
β_l	0.631	0.592	0.669	0.198
β_k	0.396	0.417	0.343	0.130
β_m	-	-	-	0.688
N	74,988	129,043	181,985	181,985
		Services		
β_l	0.618	0.598	0.681	0.446
β_k	0.402	0.339	0.384	0.215
β_m	-	-	-	0.354
N	77,209	146,717	173,332	173,332

Table 10 contains returns to scale estimates by industry, at the 2-digit SIC level. The number of firms on which estimation was computed is included. If the factor elasticity on labour, capital, or materials was outside the range of $[0,1]$, then the RTS was not computed.

SIC	ACF	GNR	N
10	1.033	1.039	12,495
11	1.187	1.092	1,724
13	1.186	1.020	4,981
14	-	0.996	3,355
15	1.019	1.115	841
16	1.244	1.057	3,478
17	1.335	1.012	4,184
18	1.156	1.054	7,521
19	1.097	-	506
20	1.418	1.016	5,733
21	1.164	1.061	986
22	1.173	1.026	7,776
23	-	1.016	5,616
24	-	1.001	4,776
25	-	1.019	15,597
26	1.085	1.036	7,648
27	1.117	1.011	4,913
28	1.289	1.023	10,899
29	-	1.181	1,633
30	1.440	1.091	1,973
31	-	1.050	4,060
32	1.350	1.055	5,020
33	1.244	1.065	4,997
41	1.102	1.067	12,216
42	0.918	1.038	12,554
43	1.330	1.059	27,014
45	1.018	1.036	24,639
46	0.756	1.041	68,969
47	1.010	1.061	66,171

SIC	ACF	GNR	N
49	1.214	1.046	11,501
50	0.972	1.094	1,306
51	1.045	1.107	807
52	1.140	1.066	8,103
53	1.270	1.429	489
55	1.517	1.011	8,549
56	1.394	0.971	25,219
58	1.128	1.032	6,802
59	1.151	1.008	2,547
60	1.292	1.086	693
61	1.062	1.131	1,062
62	-	1.103	9,061
63	1.066	1.119	1,224
69	0.538	1.045	10,295
70	1.004	1.053	10,274
71	-	1.032	11,953
72	0.888	1.022	2,323
73	0.944	1.054	5,168
74	1.010	1.079	4,769
75	1.315	0.987	1,482
77	1.006	1.041	6,195
78	1.090	1.010	9,842
79	1.145	1.094	4,136
80	1.076	1.072	1,926
81	1.143	1.042	6,472
82	0.975	1.109	9,624
90	0.846	0.936	3,111
91	-	-	1,722
92	1.127	1.030	1,248
93	1.066	1.025	7,853
94	1.264	1.045	6,086
95	-	1.117	1,889
96	1.064	0.979	11,807

Table 10: Estimates of returns to scale across 2-digit SICs, following the Ackerberg, Caves, and Frazer (2015) and Gandhi, Navarro, and Rivers (2020) approaches with a Cobb-Douglas production function. Missing sectors have estimated coefficients on labour, capital, or materials that are negative or greater than one.

Table 11: Changing Elasticity Estimates, 1998 - 2014.

	1998 - 2001	2002 - 2005	2006 - 2009	2010 - 2014
		Economy-Wide		
β_l	0.422	0.608	0.656	0.719
β_k	0.566	0.474	0.389	0.342
N	153,874	144,465	108,619	120,855
		Manufacturing		
β_l	0.577	0.743	0.743	0.773
β_k	0.537	0.498	0.353	0.375
N	41,572	36,074	24,280	21,626
		Construction		
β_l	0.706	0.633	0.845	0.673
β_k	0.205	0.238	0.234	0.621
N	13,050	13,180	9,797	14,145
		Wholesale/Trade/Transport		
β_l	0.680	0.612	0.623	0.730
β_k	0.449	0.432	0.376	0.441
N	32,792	31,360	27,476	37,415
		Services		
β_l	0.583	0.607	0.637	0.809
β_k	0.434	0.419	0.382	0.300
N	34,698	34,241	32,070	45,708

All estimates follow Ackerberg, Caves, and Frazer (2015) with a value-added Cobb-Douglas production function.

Table 12: Changing Elasticity Estimates, 1998 - 2014.

	1998 - 2001	2002 - 2005	2006 - 2009	2010 - 2014
		Economy-Wide		
β_l	0.265	0.313	0.367	0.390
β_k	0.121	0.106	0.194	0.245
β_m	0.609	0.610	0.471	0.391
N	153,874	144,465	108,619	120,855
		Manufacturing		
β_l	0.265	0.313	0.367	0.390
β_k	0.121	0.106	0.194	0.245
β_m	0.609	0.610	0.471	0.391
N	39,876	34,678	24,011	27,070
		Construction		
β_l	0.249	0.245	0.444	0.391
β_k	0.149	0.123	0.205	0.306
β_m	0.619	0.644	0.443	0.365
N	13,484	13,416	10,210	13,269
		Wholesale/Trade/Transport		
β_l	0.148	0.173	0.222	0.294
β_k	0.073	0.088	0.194	0.217
β_m	0.788	0.813	0.619	0.542
N	53,814	50,631	37,906	40,965
		Services		
β_l	0.388	0.457	0.482	0.445
β_k	0.183	0.132	0.234	0.290
β_m	0.412	0.421	0.327	0.271
N	46,700	45,740	36,492	39,551

All estimates follow Gandhi, Navarro, and Rivers (2020) with a gross-output Cobb-Douglas production function.

D Returns to Scale and Fixed Costs

We proxy average fixed costs at the 2-digit SIC industry level with *disposal of buildings, vehicles, and other assets*. We show in Figure 9 that there is no correlation between returns to scale and fixed costs across industries. This matches our theoretical prediction.

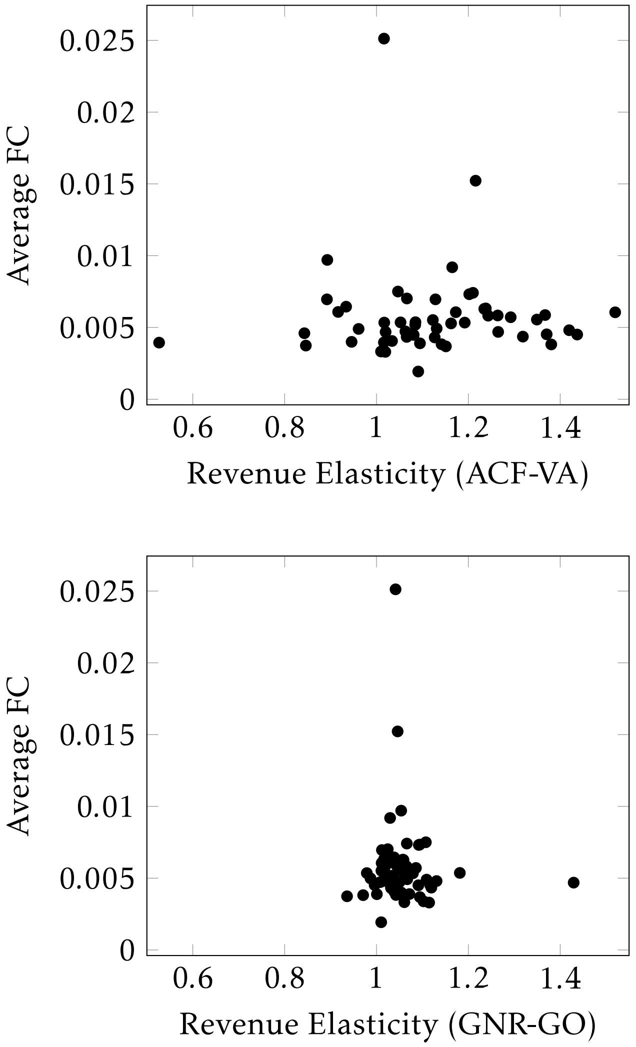


Figure 9: Relationship between Revenue Elasticity and Fixed Costs

E Model TFPR Decomposition

We follow Decker, J. Haltiwanger, Jarmin, and Miranda (2020, p. 3, 961) and classify our revenue function residual ($TFPR_t(j)$) in equation (56) as a composite of two shocks: true technical efficiency and a demand shifter. To observe this, note that revenue is the product of the inverse demand function $p_t(j) = N_t^{\frac{1+\epsilon-\mu}{\mu}} \left(\frac{Y_t}{y_t(j)}\right)^{\frac{\mu-1}{\mu}}$ and the production function $y_t(j) = A_t(j)^{1-\nu}\left[k_t(j)^{\alpha}\ell_t(j)^{1-\alpha}\right]^{\nu}$, thus:

$$p_t(j)y_t(j) = N_t^{\frac{1+\epsilon-\mu}{\mu}} Y_t^{\frac{\mu-1}{\mu}} y_t(j)^{\frac{1}{\mu}} = N_t^{\frac{1+\epsilon-\mu}{\mu}} Y_t^{\frac{\mu-1}{\mu}} A_t(j)^{\frac{1-\nu}{\mu}} \left[k_t(j)^{\alpha}\ell_t(j)^{1-\alpha}\right]^{\frac{\nu}{\mu}}. \tag{70}$$

Taking logs yields:

$$\ln p_t(j)y_t(j) = \alpha\frac{\nu}{\mu}\ln k_t(j) + (1-\alpha)\frac{\nu}{\mu}\ln \ell_t(j) + \frac{1-\nu}{\mu}\ln A_t(j) + \frac{\mu-1}{\mu}\ln Y_t + \frac{1+\epsilon-\mu}{\mu}\ln N_t. \tag{71}$$

Therefore the revenue function residual is:

$$\ln TFPR_t(j) = \frac{\nu}{\mu}(1-\alpha)\left[\ln \ell_t(j) - \frac{\ell_t^{\text{tot}}(j)}{\ell_t(j)}\ln \ell_t^{\text{tot}}(j)\right] + \frac{1-\nu}{\mu}\ln A_t(j) + \frac{\mu-1}{\mu}\ln Y_t + \frac{1+\epsilon-\mu}{\mu}\ln N_t. \tag{72}$$

This equation parallels Decker, J. Haltiwanger, Jarmin, and Miranda (2020, eq. (3)) given the following assumption. We assume that ϕ is small such that the first term is negligible. In the model ϕ ensures that some entering firms are inactive, but it cannot be too large or a high portion of entering firms will be inactive. In the data ϕ is an elusive concept: it is a measurable amount of labour (number of workers) that is set aside each period by all firms in order to begin production.[13] The $\frac{1-\nu}{\mu}\ln A_t(j)$ term represents true firm-level technical efficiency. The $\frac{\mu-1}{\mu}\ln Y_t + \frac{1+\epsilon-\mu}{\mu}\ln N_t$ term represents a demand shock. This is demand $p_t(j)$ once it is purged of individual effects. Hence, it is the logarithm of $p_t(j)y_t(j)^{\frac{\mu-1}{\mu}} = N_t^{\frac{1+\epsilon-\mu}{\mu}} Y_t^{\frac{\mu-1}{\mu}}$.

Under a standard CES aggregator, external returns to scale (love of variety) are

[13] Assuming ϕ is small is equivalent to estimating the revenue function with production labour which yields revenue elasticity of ν/μ. This is a lower bound on total revenue elasticity that is associated with very large firms.

$1+\epsilon=\mu$. In this case N_t disappears from $\ln TFPR_t(j)$. Therefore, the demand shifter is equal to $\frac{\mu-1}{\mu}\ln Y_t$, or scaled industry output. Under a no external returns to scale assumption, $\epsilon=0$, the demand shifter is scaled average output per firm $\frac{\mu-1}{\mu}\ln\frac{Y_t}{N_t}$.

Equation (61) provides a decomposition of $\ln T\bar{F}PR$ into three components. The first component $\frac{\nu}{\mu}(1-\alpha)\left[\ln\bar{\ell}-\left(1+\frac{\phi}{\bar{\ell}}\right)\ln\left(\bar{\ell}+\phi\right)\right]$ is a function of exogenous parameters. The second component $\frac{1-\nu}{\mu}\ln A(\bar{j})$ represents true average firm technical efficiency. The final term $\frac{\mu-1}{\mu}\ln Y_t+\frac{1+\epsilon-\mu}{\mu}\ln N_t$ is a demand shock. Each of these components varies by industry as ν changes.

Figure 10 plots the contribution of each component to $T\bar{F}PR$ respectively for ν from 0.75 to 0.90. It's clear that $T\bar{F}PR$ is declining in ν, so even the inclusion of terms unrelated to true average productivity $\bar{A}$ does not overturn our key result. The vast majority of the variation in $T\bar{F}PR$ is driven by the changes to $\bar{A}$. Figure 11 presents the percentage contribution of each component to $T\bar{F}PR$ for different values of ν. While $\bar{A}$ contributes over 80% for $\nu<0.76$, it makes up under half of $T\bar{F}PR$ when $\nu=0.90$.

Figure 10: Contribution of each component to $T\bar{F}PR$ for range of ν's.

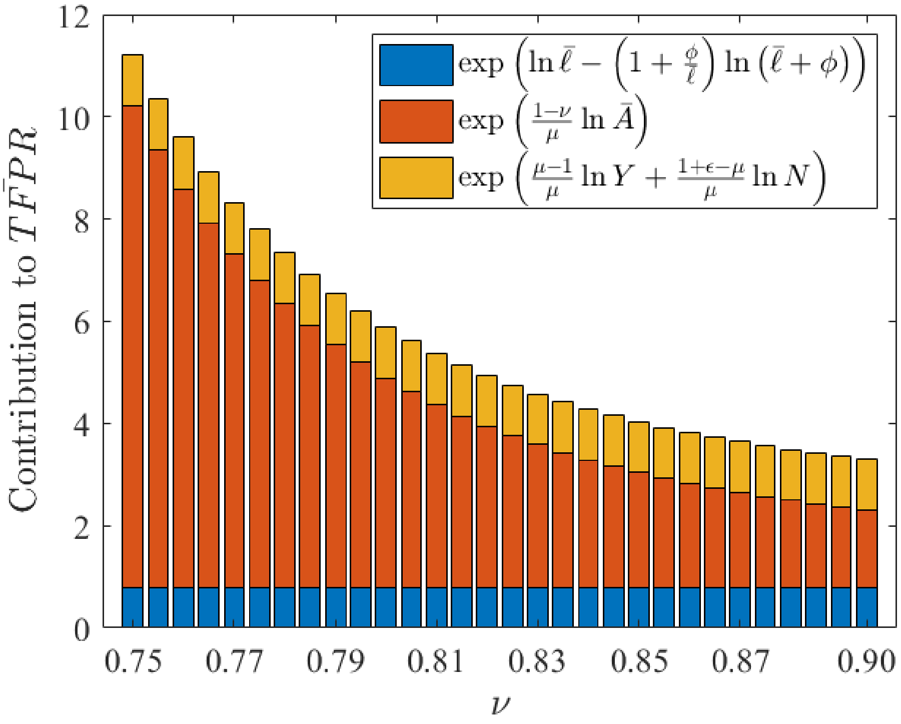
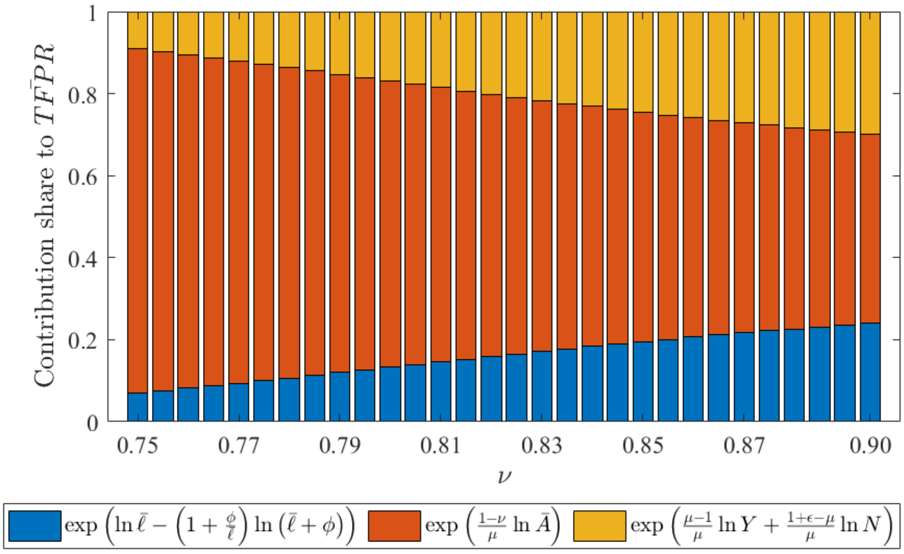


Figure 11: Percentage contribution of each component to $T\bar{F}PR$ for range of ν's.



F Productivity Estimates

Figure 12 shows the steady rise in productivity up to 2008, followed by the plateauing from the late 2000s. The estimate provided follows Gandhi, Navarro, and Rivers (2020). This result is robust across estimation methods (Ackerberg, Caves, and Frazer 2015), although the levels differ slightly (see Figure 7 in the Appendix).

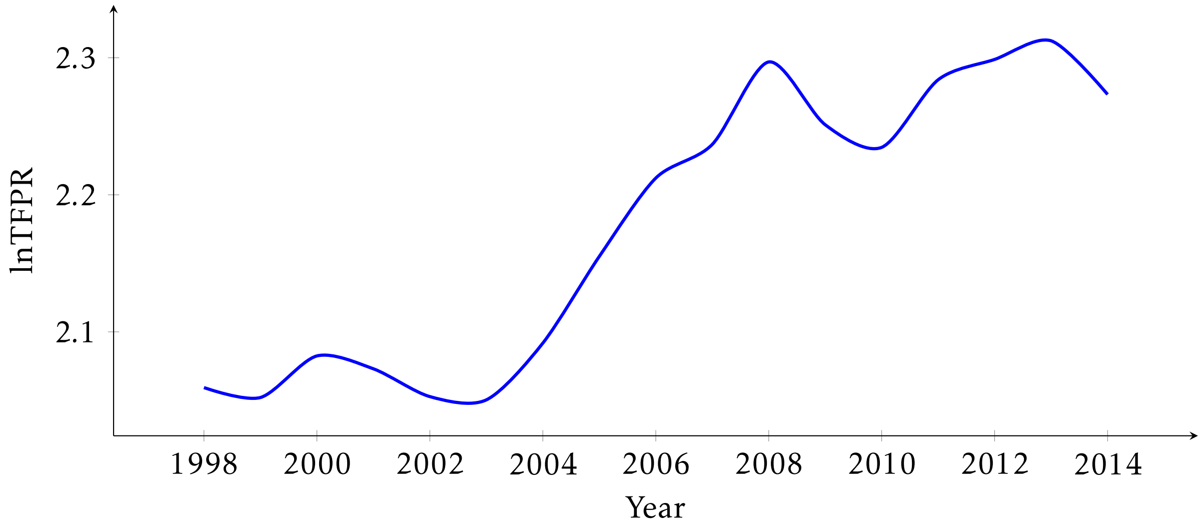


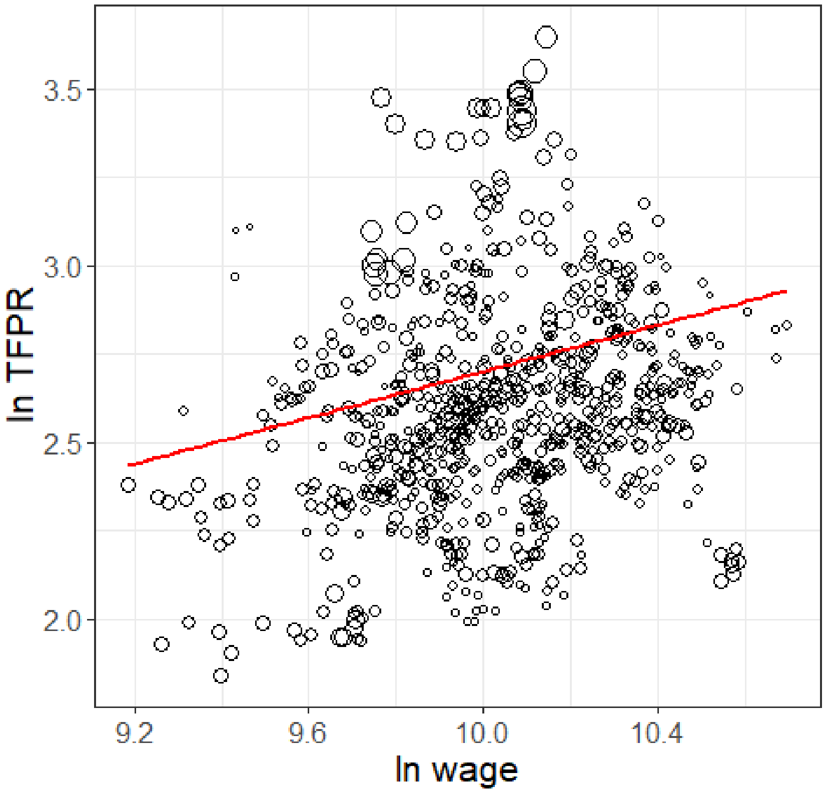
Figure 12: Aggregate TFPR (GNR)

G Wages and TFPR

Section 2.5 describes the general equilibrium effect which underpins the relationship between variable returns to production ν and the productivity cut-off J_t. It requires a rise in ν to put downwards pressure on the equilibrium wage. Therefore, the model predicts a positive relationship between the equilibrium wage and average productivity in a sector.

We obtain data on median annual full-time wages by 2-digit SIC from the Annual Survey of Hours and Earnings.[14] This is combined with our annual $\ln T\bar{F}PR$ estimates at the industry level, computed with a value-added Cobb-Douglas production function following Ackerberg, Caves, and Frazer (2015). Figure 13 presents a scatter plot of each year × 2-digit SIC for $\ln T\bar{F}PR$ against ln wage. The line of best fit is weighted by the number of firms in each observation.[15]

Figure 13: Scatter of annual log wage and log TFPR by 2-digit SIC.



[14] https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/regionbyindustry2digitsicashetable5

[15] A weighted log-log regression of $T\bar{F}PR$ on the wage with year fixed-effects yields a coefficient of 0.319 with a year-clustered standard error of 0.084.

H Further Figures & Tables

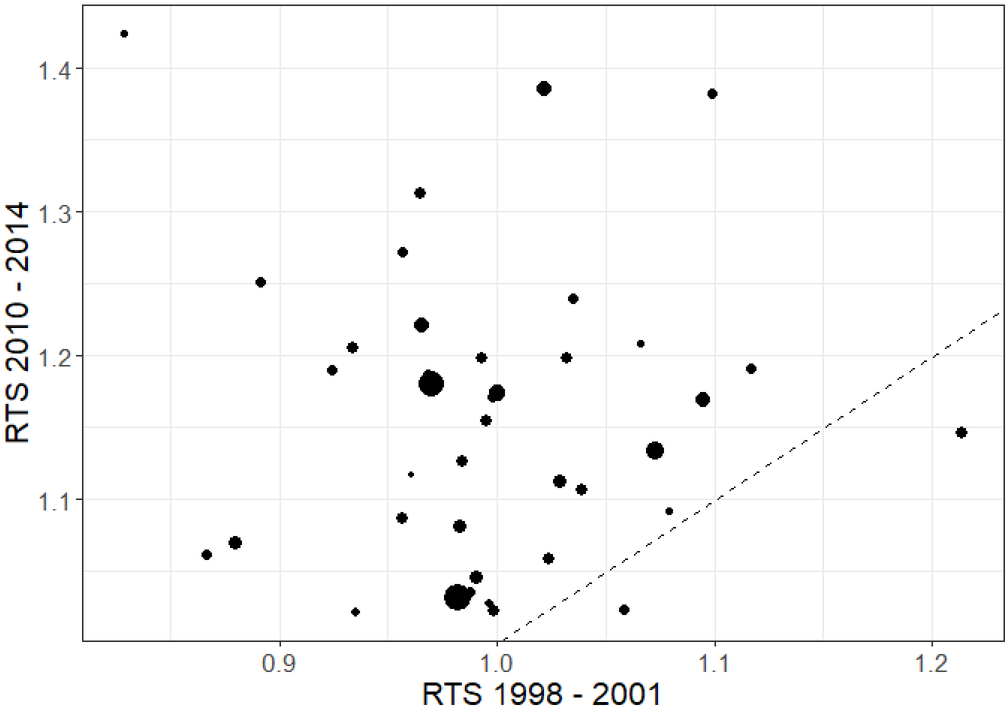
Table 13 compares the share of firms and employment within five bins of firm size, both in the UK data and the calibrated model. For example, the first row shows that around 66% of firms are non-employers (i.e. they have one self-employed worker), comprising almost 40% of total employment. By contrast, the model has almost 75% of such firms, employing 17% of workers.

Table 13: UK Firm Distribution: Data & Model

Firm Size	Firm Share		Employment Share	
	Model	Data	Model	Data
1	0.7467	0.6571	0.1683	0.3926
2 - 9	0.2079	0.2275	0.1561	0.2457
10 - 49	0.0377	0.0577	0.1524	0.1031
50 - 249	0.0064	0.0354	0.1288	0.0979
250+	0.0014	0.0223	0.3944	0.1607

Data from the ONS: https://www.gov.uk/government/statistics/business-population-estimates-2021/business-population-estimates-for-the-uk-and-regions-2021-statistical-release-html.

Figure 14: Changing Returns to Scale by 2-digit SIC, GNR Estimation.



*

Figure 15: Comparison of returns to scale at 2-digit SIC level, from 1998 - 2001 to 2010 - 2014. Size of points represents the average number of firms in that sector in each period. Dotted line is 45 degree line: points above that line are consistent with a rise in returns to scale.

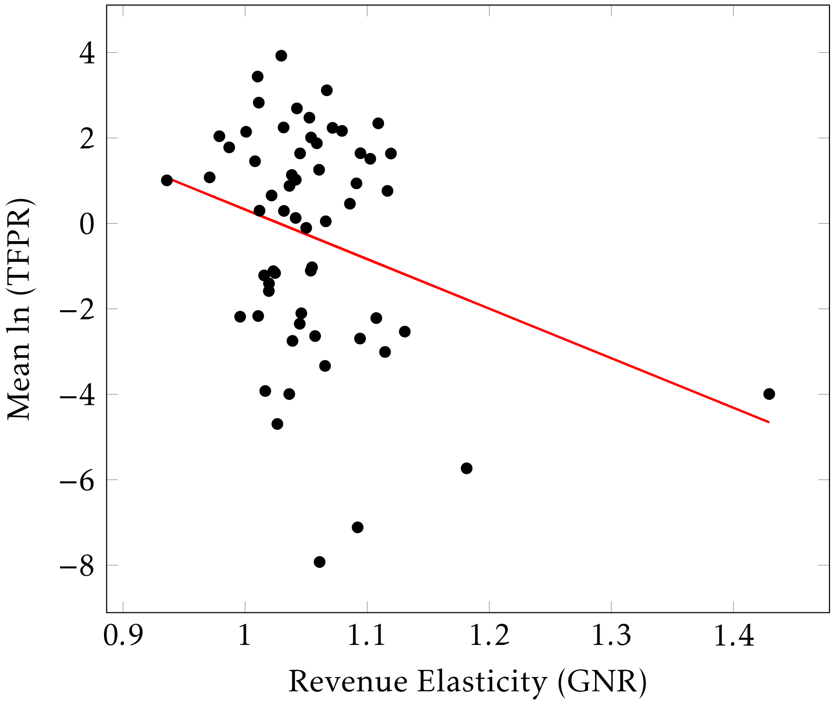


Figure 16: Relationship between Revenue Elasticity and TFP, GNR

I Productivity Distribution

We obtain a measure of productivity $A(\jmath)$ from a random draw on the unit interval $\jmath \in [0,1]$ using inverse transform sampling. For example, we could model the distribution of firm productivity (or size) with a Pareto distribution or log-normal distribution (Axtell 2001; Atkeson and P. J. Kehoe 2005; Luttmer 2007; Barseghyan and DiCecio 2011; Asturias, Hur, T. J. Kehoe, and Ruhl 2022). The Pareto CDF is given by

$$F(A;\vartheta) = 1 - \left(\frac{h}{A}\right)^{\vartheta}; \quad A \geq h > 0 \quad \text{and} \quad \vartheta > 0.$$

If $\mathcal{J} \sim Uniform(0,1]$, then for $\jmath \in \mathcal{J}$, we have

$$1 - \left(\frac{h}{A}\right)^{\vartheta} = \jmath$$

Therefore

$$A(\jmath) = h(1-\jmath)^{-\frac{1}{\vartheta}}.$$

Typically we set the scale parameter, which is the minimum possible value of A, to $h = 1$. Calibrations of the shape parameter (tail index) vary, for example $\vartheta = 1.15$ in Barseghyan and DiCecio (2011) and $\vartheta = 1.06$ in Luttmer (2007) and $\vartheta = 6.10$ in Asturias, Hur, T. J. Kehoe, and Ruhl (2022). These estimates are set to match the firm size distribution in terms of employment since in the model productivity is roughly proportional to employment.

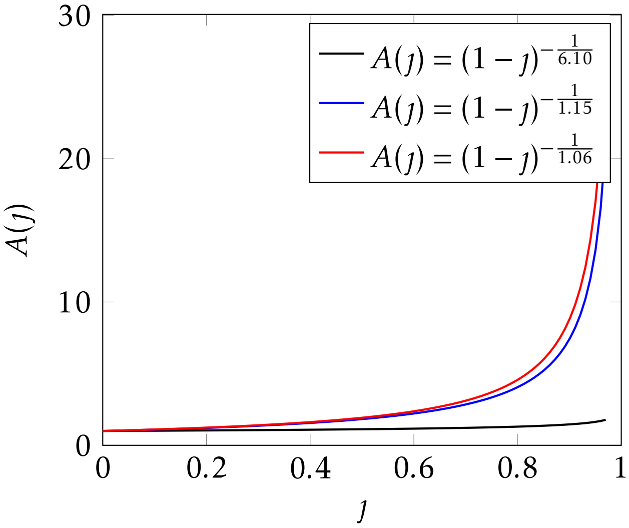


Figure 17: Productivity with Pareto Distribution, $h = 1, \vartheta = \{1.06, 1.15\}$

J Mapping Model to Data with Output not Revenue

There are two nuances to our estimation process. The first is that we observe total labour not production labour and the second is that we observe revenue not output. To reduce the number of moving parts, consider that we we observe output, rather than revenue, but still we cannot estimate the firm-level production function (9) directly because we do not have production labour. In this case our estimating equation is:

$$\ln y_t(\jmath) = (1-\nu)\ln A_t(\jmath) + \beta_1 \ln k_t(\jmath) + \beta_2 \ln \ell_t^{\text{tot}}(\jmath) \tag{73}$$

The coefficients β_1 and β_2 represent the elasticity of firm-level output to firm-level capital and firm-level total labour, respectively. The coefficients are treated as common across all $\jmath = 1 \ldots N$ within each industry and common across time. The sum of the coefficient is the response of output to a change in *all* inputs. The coefficients are

$$\beta_1 = \frac{\partial \ln y_t(\jmath)}{\partial \ln k_t(\jmath)} = \nu\alpha \tag{74}$$

$$\beta_2 = \frac{\partial \ln y_t(\jmath)}{\partial \ln \ell_t^{\text{tot}}(\jmath)} = \nu(1-\alpha)\left(1 + \frac{\phi}{\ell_t(\jmath)}\right) \tag{75}$$

The first equality follows from taking the derivatives of (73) and the second equality follows from taking the derivative of the production function multiplied by inverse demand, we specify in logs below:

$$\ln y_t(\jmath) = (1-\nu)\ln A_t(\jmath) + \alpha\nu \ln k_t(\jmath) + (1-\alpha)\nu \ln\left[\ell_t^{\text{tot}}(\jmath) - \phi\right]. \tag{76}$$

Therefore the sum of the regression coefficients give returns to scale (RTS) (the change in output from a change in all inputs):

$$\beta_1 + \beta_2 = \nu\left(1 + (1-\alpha)\frac{\phi}{\ell_t(\jmath)}\right) \tag{77}$$

The coefficients are for the average firm in an industry and average over the time period, therefore the $\phi/\ell_t(\jmath)$ is constant representing the overhead to production labour ratio for the average firm in the average year.

K Additional Derivations

K.1 Household Utility Maximization Problem

$$\mathcal{L} = \sum_{t=0}^{\infty} \beta^t \left\{ \frac{C_t^{1-\sigma} - 1}{1-\sigma} + \lambda\left[w_t L^s + r_t K_t + \Pi_t + T_t - C_t - K_{t+1} + (1-\delta)K_t\right] \right\} \tag{78}$$

The optimality conditions are

$$\frac{1}{C_t^\sigma} = \lambda \tag{79}$$

$$\lambda = \beta\lambda_{t+1}(r_{t+1} + (1-\delta)) \tag{80}$$

$$C_t + K_{t+1} - (1-\delta)K_t = r_t K_t + w_t L^s + \Pi_t + T_t \tag{81}$$

K.2 Final Goods Producer Problem

It helps to use integration by substitution to re-define the limits such that the set of operating firms $\imath \in (0, N_t)$ is indexed on the unit interval $s \in (0,1)$. Thus, $s = \imath / N_t$, hence $N_t ds = d\imath$ and the problem is:

$$\Pi^F = \max_{y_t(s)} \quad Y_t - N_t \int_0^1 p_t(s) y_t(s) ds \qquad \text{s.t.} \quad Y_t = N_t^{1+\epsilon} \left[\int_0^1 y_t(s)^{\frac{1}{\mu}} ds \right]^{\mu}.$$

This yields the inverse-demand:

$$p_t(s) = N_t^{\frac{1+\epsilon-\mu}{\mu}} \left(\frac{y_t(s)}{Y_t} \right)^{\frac{1-\mu}{\mu}}.$$

K.3 Firm Profit Maximization Problem (one-stage)

$$\begin{aligned} \mathcal{L}(k, \ell, y, \lambda) &= p_t(\jmath) y_t(\jmath) - r_t k_t(\jmath) - w_t(\ell_t(\jmath) + \phi) \\ &+ \lambda(\jmath) \left[A_t^{1-\nu} \left(k_t(\jmath)^{\alpha} \ell_t(\jmath)^{1-\alpha} \right)^{\nu} - y_t(\jmath) \right] \\ &+ \Delta(\jmath) \left(N_t^{\frac{1+\epsilon-\mu}{\mu}} \left(\frac{y_t(\jmath)}{Y_t} \right)^{\frac{1-\mu}{\mu}} - p_t(\jmath) \right) \end{aligned} \tag{82}$$

where $\lambda(\jmath)$ and $\Delta(\jmath)$ are Lagrange multipliers. The corresponding optimization conditions are

$$\begin{aligned}
\ell_t(\jmath): &\quad w_t - \lambda(\jmath)\nu(1-\alpha)\frac{y_t(\jmath)}{\ell_t(\jmath)} = 0 &\therefore&\quad w_t = \lambda(\jmath)\nu(1-\alpha)\frac{y_t(\jmath)}{\ell_t(\jmath)} &(83)\\
k(\jmath): &\quad r_t - \lambda(\jmath)\nu\alpha\frac{y_t(\jmath)}{k_t(\jmath)} = 0 &\therefore&\quad r_t = \lambda(\jmath)\nu\alpha\frac{y_t(\jmath)}{k_t(\jmath)} &(84)\\
y(\jmath): &\quad p_t(\jmath) - \lambda(\jmath) + \Delta(\jmath)\left(\frac{1-\mu}{\mu}\right)\frac{p_t(\jmath)}{y_t(\jmath)} = 0 &\therefore&\quad p_t(\jmath) = \lambda(\jmath)\left(1 + \frac{\Delta(\jmath)}{y_t(\jmath)}\left(\frac{1-\mu}{\mu}\right)\right)^{-1} &(85)\\
p(\jmath): &\quad y_t(\jmath) - \Delta_t(\jmath) = 0 &\therefore&\quad y_t(\jmath) = \Delta(\jmath) &(86)\\
\lambda(\jmath): &\quad y_t(\jmath) - A_t^{1-\nu}\left(k_t(\jmath)^{\alpha}\ell_t(\jmath)^{1-\alpha}\right)^{\nu} = 0 &\therefore&\quad y_t(\jmath) = A_t^{1-\nu}\left(k_t(\jmath)^{\alpha}\ell_t(\jmath)^{1-\alpha}\right)^{\nu} &(87)\\
\Delta(\jmath): &\quad N_t^{\frac{1+\epsilon-\mu}{\mu}}\left(\frac{y_t(\jmath)}{Y_t}\right)^{\frac{1-\mu}{\mu}} - p_t(\jmath) = 0 &\therefore&\quad p_t(\jmath) = N_t^{\frac{1+\epsilon-\mu}{\mu}}\left(\frac{y_t(\jmath)}{Y_t}\right)^{\frac{1-\mu}{\mu}} &(88)
\end{aligned}$$

K.4 Two-Stage Cost-Min Profit-Max

A two-stage approach, which separates the cost minimization and profit maximization problem, helps us to stress which results we can derive in the absence of a specific demand system.

K.4.1 Cost Minimisation

The cost minimization is as follows where prices are in nominal terms. The full solution is found in Appendix K.5. The overhead labour cost is treated as independent of optimal size of production decisions, except in determining whether to produce or not. This is because it is a one-off cost that must be paid regardless of the size of operation.

$$\mathcal{C}(\jmath) = \min_{\ell_t(\jmath), k_t(\jmath)} \quad w_t(\ell_t(\jmath) + \phi) + r_t k_t(\jmath), \tag{89}$$

$$\text{s.t.} \quad y_t(\jmath) \leq A_t^{1-\nu} \left(k_t(\jmath)^{\alpha} \ell_t(\jmath)^{1-\alpha} \right)^{\nu}. \tag{90}$$

The optimality conditions from the firm cost minimization problem imply factor expenditures are given by

$$w_t \ell_t(r_t, w_t, y(\jmath)) = (1-\alpha) \left(\frac{w_t}{1-\alpha} \right)^{1-\alpha} \left(\frac{r_t}{\alpha} \right)^{\alpha} \left(\frac{y_t(\jmath)}{A(\jmath)^{1-\nu}} \right)^{\frac{1}{\nu}} \tag{91}$$

$$r_t k_t(r_t, w_t, y(\jmath)) = \alpha \left(\frac{w_t}{1-\alpha} \right)^{1-\alpha} \left(\frac{r_t}{\alpha} \right)^{\alpha} \left(\frac{y_t(\jmath)}{A(\jmath)^{1-\nu}} \right)^{\frac{1}{\nu}}. \tag{92}$$

Therefore total costs are given by

$$\mathcal{C}(r_t, w_t, y_t(\jmath)) = \left(\frac{w_t}{1-\alpha} \right)^{1-\alpha} \left(\frac{r_t}{\alpha} \right)^{\alpha} \left(\frac{y_t(\jmath)}{A(\jmath)^{1-\nu}} \right)^{\frac{1}{\nu}} + w_t \phi. \tag{93}$$

Marginal cost is given by

$$\frac{d\mathcal{C}}{dy} = \frac{1}{\nu} \frac{\mathcal{C} + w_t \phi}{y} = \frac{1}{\nu} \left(\frac{w_t}{1-\alpha} \right)^{1-\alpha} \left(\frac{r_t}{\alpha} \right)^{\alpha} \left(\frac{y_t(\jmath)}{A(\jmath)} \right)^{\frac{1-\nu}{\nu}}. \tag{94}$$

The elasticity of marginal cost to output is constant

$$\frac{d\ln\left(\frac{d\mathcal{C}}{dy}\right)}{d\ln y} = \frac{1-\nu}{\nu}. \tag{95}$$

K.4.2 Profit Maximization

The profit maximization problem of the intermediate goods producer is as follows where the cost function is treated as given

$$\pi_t(\jmath) = \max_{y_t(\jmath)} p_t(\jmath)y_t(\jmath) - \mathcal{C}(r_t, w_t, y_t(\jmath)) \tag{96}$$

$$\text{s.t.} \quad p_t(\jmath) = N_t^{\frac{1+\epsilon-\mu}{\mu}} \left(\frac{y_t(\jmath)}{Y_t}\right)^{\frac{1-\mu}{\mu}}. \tag{97}$$

The first-order condition for profit maximization is

$$\frac{p_t(\jmath)}{\frac{\partial\mathcal{C}(r_t, w_t, y_t(\jmath))}{\partial y_t(\jmath)}} = \mu \tag{98}$$

The first-order condition states that a firm chooses output such that that price is a constant markup above marginal cost.

K.5 Lagrange Solution to Cost Minimization Problem

The Lagrangean for the cost minimization problem is given by

$$\mathcal{L}(k,\ell,\lambda) = w_t(\ell_t(\jmath) + \phi) + r_t k_t(\jmath) - \lambda(\jmath)\left[A_t^{1-\nu}\left(k_t(\jmath)^\alpha \ell_t(\jmath)^{1-\alpha}\right)^\nu - y_t(\jmath)\right] \tag{99}$$

where $\lambda(\jmath)$ is the Lagrange multiplier. The corresponding optimization conditions are

$$\ell_t(\jmath): \qquad w_t - \lambda(\jmath)\nu(1-\alpha)\frac{y_t(\jmath)}{\ell_t(\jmath)} = 0 \qquad \therefore \quad w_t = \lambda(\jmath)\nu(1-\alpha)\frac{y_t(\jmath)}{\ell_t(\jmath)} \tag{100}$$

$$k(\jmath): \qquad r_t - \lambda(\jmath)\nu\alpha\frac{y_t(\jmath)}{k_t(\jmath)} = 0 \qquad \therefore \quad r_t = \lambda(\jmath)\nu\alpha\frac{y_t(\jmath)}{k_t(\jmath)} \tag{101}$$

$$\lambda(\jmath): \qquad y_t(\jmath) - A_t^{1-\nu}\left(k_t(\jmath)^\alpha \ell_t(\jmath)^{1-\alpha}\right)^\nu = 0 \quad \therefore \quad y_t(\jmath) = A_t^{1-\nu}\left(k_t(\jmath)^\alpha \ell_t(\jmath)^{1-\alpha}\right)^\nu \tag{102}$$

The three conditions combine to yield

$$y_t(\jmath) = A(\jmath)\left[\lambda(\jmath)\nu\left(\frac{\alpha}{r_t}\right)^{\alpha}\left(\frac{1-\alpha}{w_t}\right)^{1-\alpha}\right]^{\frac{\nu}{1-\nu}}, \tag{103}$$

$$\therefore \lambda(\jmath) = \frac{1}{\nu}\left(\frac{r_t}{\alpha}\right)^{\alpha}\left(\frac{w_t}{1-\alpha}\right)^{1-\alpha}\left(\frac{y_t(\jmath)}{A(\jmath)}\right)^{\frac{1-\nu}{\nu}}. \tag{104}$$

The optimality conditions from the firm cost minimization problem imply that the ratio of capital to labour (capital intensity) is the same at all firms:

$$\frac{k_t(\jmath)}{\ell_t(\jmath)} = \frac{w_t}{r_t}\frac{1-\alpha}{\alpha} \qquad \forall \jmath \in (0,1). \tag{105}$$

The result arises because all firms operate with the same technology, which is defined by α, ν, and there are perfect factor markets, meaning that all firms pay the same price r_t, w_t for capital and labour. Combining the ratio of cost-minimizing input demand with output yields optimal factor demands as a function of output and prices

$$\ell_t(r_t, w_t, y(\jmath)) = \left(\frac{r_t}{w_t}\frac{1-\alpha}{\alpha}\right)^{\alpha}\left(\frac{y_t(\jmath)}{A(\jmath)^{1-\nu}}\right)^{\frac{1}{\nu}} \tag{106}$$

$$k_t(r_t, w_t, y(\jmath)) = \left(\frac{w_t}{r_t}\frac{\alpha}{1-\alpha}\right)^{1-\alpha}\left(\frac{y_t(\jmath)}{A(\jmath)^{1-\nu}}\right)^{\frac{1}{\nu}}. \tag{107}$$

Given the optimality conditions for capital and labour ((100) and (101)), the cost function is given by

$$\mathcal{C}(\jmath) = w_t\ell_t(\jmath) + r_t k_t(\jmath) + w_t\phi = \lambda(\jmath)\nu\left(1 + (1-\alpha)\frac{\phi}{\ell_t(\jmath)}\right)y_t(\jmath). \tag{108}$$

At this point, it is helpful to note that the Lagrange multiplier $\lambda(\jmath)$ of the constraint is equal to the rate of change of the solution to the constrained minimization problem as the constraint varies, in other words the marginal cost:

$$\frac{d\mathcal{C}(y_t(\jmath))}{dy_t(\jmath)} = \lambda(\jmath; y_t(\jmath)). \tag{109}$$

Since returns to scale are the ratio of average cost to marginal costs. Divide equation (108) by output $y_t(\jmath)$ to get average cost, and by marginal cost $\lambda(\jmath)$ to get returns to scale

$$\mathrm{RTS}_t(\jmath) = \nu\left(1 + (1-\alpha)\frac{\phi}{\ell_t(\jmath)}\right). \tag{110}$$

Crucially the degree of returns to scale depends on the ratio of fixed labour overheads ϕ to production labour $\ell_t(\jmath)$.

K.6 Factor Shares

The share of production labour in revenue and the share of capital rental payments in revenue are

$$\frac{w_t\ell_t(\jmath)}{p_t(\jmath)y_t(\jmath)} = \frac{\lambda(\jmath)}{p_t(\jmath)}(1-\alpha)\nu = \frac{(1-\alpha)\nu}{\mu} \tag{111}$$

$$\frac{r_tk_t(\jmath)}{p_t(\jmath)y_t(\jmath)} = \frac{\lambda(\jmath)}{p_t(\jmath)}\alpha\nu = \frac{\alpha\nu}{\mu}. \tag{112}$$

The first equality relies on cost minimization only. The second equality uses the profit maximization results.

K.7 Variable Cost Share

Total variable costs at factor market equilibrium are given by

$$w_t\ell_t(\jmath) + r_tk_t(\jmath) = \frac{\nu}{\mu}p_t(\jmath)y_t(\jmath). \tag{113}$$

The ratio ν/μ is the variable cost share in revenue which is also the gross profit share residual where gross profits refer to profits before fixed costs are deducted:

$$\frac{\nu}{\mu} = \frac{w_t\ell_t(\jmath)}{p_t(\jmath)y_t(\jmath)} + \frac{r_tk_t(\jmath)}{p_t(\jmath)y_t(\jmath)} = 1 - \left(\frac{\pi(\jmath)}{p(\jmath)y(\jmath)} + \frac{w\phi}{p(\jmath)y(\jmath)}\right). \tag{114}$$

The ratio is often calibrated to 0.85 which implies the gross profit share is 15% or the sum of capital and (production) labour share is 85%. Hopenhayn (2014) provides a discussion of this calibration.

K.8 Relative Productivity

By taking the ratio of two firms in the same factor market we can show that relative input choices correspond to relative size. Consider two different firms $\imath \neq \jmath$ they will face the same rental on capital and wage:

$$r_t = \alpha \frac{\nu}{\mu} \frac{p_t(\jmath) y_t(\jmath)}{k_t(\jmath)} = \alpha \frac{\nu}{\mu} \frac{p_t(\imath) y_t(\imath)}{k_t(\imath)} \tag{115}$$

$$w_t = (1-\alpha) \frac{\nu}{\mu} \frac{p_t(\jmath) y_t(\jmath)}{\ell_t(\jmath)} = (1-\alpha) \frac{\nu}{\mu} \frac{p_t(\imath) y_t(\imath)}{\ell_t(\imath)} \quad \forall \imath, \jmath \in [0, N_t], \quad \imath \neq \jmath. \tag{116}$$

Hence taking the ratio for any two firms $\imath, \jmath$ means factor prices cancel-out yielding

$$\frac{p_t(\jmath) y_t(\jmath)}{p_t(\imath) y_t(\imath)} = \frac{k(\jmath)}{k(\imath)} = \frac{\ell(\jmath)}{\ell(\imath)}. \tag{117}$$

The final step is to show equivalence to scaled productivity. Notice that the ratio of the inverse-demand curve is

$$\frac{p_t(\jmath)}{p_t(\imath)} = \frac{N_t^{\frac{1+\epsilon-\mu}{\mu}} \left[\frac{y_t(\jmath)}{Y_t}\right]^{\frac{1-\mu}{\mu}}}{N_t^{\frac{1+\epsilon-\mu}{\mu}} \left[\frac{y_t(\imath)}{Y_t}\right]^{\frac{1-\mu}{\mu}}} = \left(\frac{y_t(\jmath)}{y_t(\imath)}\right)^{\frac{1-\mu}{\mu}}. \tag{118}$$

Therefore the ratio of firm revenues is

$$\frac{p_t(\jmath) y_t(\jmath)}{p_t(\imath) y_t(\imath)} = \left[\frac{y_t(\jmath)}{y_t(\imath)}\right]^{\frac{1}{\mu}} = \left[\frac{A(\jmath)^{1-\nu} \left(k(\jmath)^{\alpha} \ell(\jmath)^{1-\alpha}\right)^{\nu}}{A(\imath)^{1-\nu} \left(k(\imath)^{\alpha} \ell(\imath)^{1-\alpha}\right)^{\nu}}\right]^{\frac{1}{\mu}}. \tag{119}$$

If we use the result from factor market equilibrium that the ratio of two firms' inputs corresponds to the ratio of their size this becomes

$$\frac{p_t(\jmath) y_t(\jmath)}{p_t(\imath) y_t(\imath)} = \left[\left(\frac{A(\jmath)}{A(\imath)}\right)^{1-\nu} \left(\frac{p_t(\jmath) y_t(\jmath)}{p_t(\imath) y_t(\imath)}\right)^{\nu}\right]^{\frac{1}{\mu}} \tag{120}$$

Finally collect terms in revenue and simplify

$$\therefore \frac{p_t(\jmath)y_t(\jmath)}{p_t(\imath)y_t(\imath)} = \left(\frac{A(\jmath)}{A(\imath)}\right)^{\frac{1-\nu}{\mu-\nu}} = \frac{a(\jmath)}{a(\imath)}. \tag{121}$$

In the final step we redefine productivity $A(\jmath)$ as scaled productivity $a(\jmath)$ as follows

$$a(\jmath) \equiv A(\jmath)^{\frac{1-\nu}{\mu-\nu}}. \tag{122}$$

K.9 Zero-profit firm

Under the optimality conditions profits are given by

$$\pi_t(\jmath) = \left(1 - \frac{\nu}{\mu}\right) p_t(\jmath)y_t(\jmath) - w_t\phi. \tag{123}$$

Imposing zero-profits yields the revenue of the threshold firm

$$p_t(J_t)y_t(J_t) = \left(1 - \frac{\nu}{\mu}\right)^{-1} \phi w_t. \tag{124}$$

The threshold firm is the smallest revenue firm. It has the highest fixed cost share and greatest returns to scale, equal to the markup as the profit share is zero:

$$\frac{\phi w_t}{p(J_t)y(J_t)} = 1 - \frac{\nu}{\mu} \tag{125}$$

$$\mathrm{RTS}(J_t) = \mu \tag{126}$$

Substitute $w_t\phi$ from the threshold firm revenue equation (124) into the optimal profit condition (123) and use the ratio of firm revenue to scaled productivity result (14) yields:

$$\begin{aligned}
\pi_t(\jmath) &= \left(1-\frac{\nu}{\mu}\right)p_t(\jmath)y_t(\jmath) - \left(1-\frac{\nu}{\mu}\right)p_t(J_t)y_t(J_t) && (127)\\
&= \left(1-\frac{\nu}{\mu}\right)p_t(J_t)y_t(J_t)\left(\frac{p_t(\jmath)y_t(\jmath)}{p_t(J_t)y_t(J_t)}-1\right) && (128)\\
&= \phi w_t\left(\frac{p_t(\jmath)y_t(\jmath)}{p_t(J_t)y_t(J_t)}-1\right) && (129)\\
&= \phi w_t\left(\frac{a(\jmath)}{a(J_t)}-1\right) && (130)
\end{aligned}$$

K.10 Free Entry

$$\mathbb{E}[v_t(\jmath)] = \int_0^1 v_t(\jmath)d\jmath = \int_0^{J_t} 0d\jmath + \int_{J_t}^1 \pi_t(\jmath)d\jmath = \int_{J_t}^1 \pi_t(\jmath)d\jmath = \kappa. \tag{131}$$

Hence from equation (17)

$$\phi w_t \int_{J_t}^1 \left(\frac{a(\jmath)}{a(J_t)}-1\right)d\jmath = \kappa$$

K.11 Firm-level returns to scale

The optimal profit equation provides a link between two expressions of returns to scale.

Combining equation (41) and equation (123), returns to scale can be rewritten as a function of firm profits:

$$\begin{aligned}
RTS_t(\jmath) &= \nu + \mu\frac{w_t\phi}{p_t(\jmath)y_t(\jmath)} && (132)\\
&= \nu + (\mu-\nu)\frac{w_t\phi}{\pi_t(\jmath)+w_t\phi} && (133)
\end{aligned}$$

It is clear that when profits are zero, returns to scale equal the markup. Furthermore, using the relationship between profits and relative scaled productivity in equation (17), we get (42).

K.11.1 Average RTS

Returns to scale of the average firm $\bar{\jmath}$ under Pareto productivity distribution is:

$$RTS(\bar{\jmath}) = \nu + (\mu - \nu)\frac{a(J_t)}{a(\bar{\jmath})} = \nu + (\mu - \nu)\left(1 - \frac{1}{\vartheta}\left(\frac{1-\nu}{\mu-\nu}\right)\right) = \mu - \frac{1-\nu}{\vartheta}. \tag{134}$$

Note that returns to scale of the average firm is equivalent to taking the arithmetic average of returns to scale:

$$\overline{\text{RTS}} = \frac{1}{1-J_t}\int_{J_t}^{1} \text{RTS}(\jmath)d\jmath = \frac{1}{1-J_t}\int_{J_t}^{1}\left[\nu + (\mu-\nu)\frac{a(J_t)}{a(\jmath)}\right]d\jmath = \nu + (\mu-\nu)\frac{a(J_t)}{a(\bar{\jmath})}. \tag{135}$$

K.12 Aggregate Factor Inputs

$$K_t = \int_0^{N_t} k_t(\imath)d\imath = E_t\int_{J_t}^{1} k_t(\jmath)d\jmath \tag{136}$$

$$= \frac{N_t}{1-J_t}\int_{J_t}^{1} k_t(\jmath)d\jmath = N_t\bar{k}(J_t) \tag{137}$$

$$L_t = \int_0^{N_t}[\ell_t(\imath)+\phi]\,d\imath = E_t\int_{J_t}^{1}[\ell_t(\jmath)+\phi]\,d\jmath \tag{138}$$

$$= E_t\int_{J_t}^{1}\ell_t(\jmath)d\jmath + E_t\int_{J_t}^{1}\phi d\jmath = E_t\int_{J_t}^{1}\ell_t(\jmath)d\jmath + E_t(1-J_t)\phi = N_t\bar{\ell}_t(J_t) + N_t\phi \tag{139}$$

K.13 Aggregate Output

There are some important steps in the result. We use integration by substitution to redefine the limits such that $j = i/\mu_t$. Therefore, $\frac{dj}{di} = \frac{1}{\mu_t}$ and $\mu_t dj = di$. And, we use the mean-value theorem for definite integrals that for some $\bar{j} \in (J,1)$ the following holds:

$$y_t\left(\bar{J}_t\right)^{\frac{1}{\mu}}(1-J) = \int_J^1 y_t(\jmath)^{\frac{1}{\mu}}d\jmath. \tag{140}$$

We simplify notation such that that $y_t\left(\bar{J}_t\right) \equiv \bar{y}$.

$$
\begin{aligned}
Y_t &= N_t^{1+\epsilon}\left[\frac{1}{N_t}\int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}} d\imath\right]^{\mu} && (141)\\
&= N_t^{1+\epsilon}\left[\frac{E_t}{N_t}\int_{J_t}^{1} y_t(\jmath)^{\frac{1}{\mu}} d\jmath\right]^{\mu} && (142)\\
&= N_t^{1+\epsilon}\left[\frac{1}{1-J_t}\int_{J_t}^{1} y_t(\jmath)^{\frac{1}{\mu}} d\jmath\right]^{\mu} && (143)\\
&= N_t^{1+\epsilon}\bar{y}_t(J_t) && (144)
\end{aligned}
$$

An alternative way to view this result is to index operating firms as $s \in (0,1)$. The index $\jmath \in (0,1)$ represents all entrants, and the subset $\jmath \in (J_t, 1)$ represents operating firms.

$$
\begin{aligned}
Y_t = N_t^{1+\epsilon}\left[\frac{1}{N_t}\int_0^{N_t} y_t(i)^{\frac{1}{\mu}} di\right]^{\mu} &= N_t^{1+\epsilon}\left[\frac{1}{N_t}\int_0^{1} y_t(s)^{\frac{1}{\mu}} N_t ds\right]^{\mu} && (145)\\
= N_t^{1+\epsilon}\left[\int_0^{1} y_t(s)^{\frac{1}{\mu}} ds\right]^{\mu} &= N_t^{1+\epsilon}\bar{y}_t(s). && (146)
\end{aligned}
$$

The variable $\bar{y}(s)$ is the average value across operating firms, which is equivalent to $\bar{y}(J_t)$. Hence both approaches yields the same interpretation that aggregate output is the average output of operating firms multiplied by the number of operating firms raised to the power of any external scale economies.

Now consider average output of operating firms

$$
\begin{aligned}
\bar{y}(J_t) &= \bar{A}^{1-\nu}[\bar{k}_t^{\alpha}\bar{\ell}_t^{1-\alpha}]^{\nu} && (147)\\
&= \bar{A}^{1-\nu}\left[\left(\frac{1}{1-J}\int_J^1 k_t(\jmath)d\jmath\right)^{\alpha}\left(\frac{1}{1-J}\int_J^1 \ell_t(\jmath)d\jmath\right)^{1-\alpha}\right]^{\nu} && (148)\\
&= \bar{A}^{1-\nu}\left[\left(\frac{1}{1-J}\frac{K_t}{\nu_t}\right)^{\alpha}\left(\frac{1}{1-J}\frac{u_tL_t}{\nu_t}\right)^{1-\alpha}\right]^{\nu} && (149)\\
&= \bar{A}^{1-\nu}\left[\frac{1}{N_t}K_t^{\alpha}\left(u_tL_t\right)^{1-\alpha}\right]^{\nu} && (150)
\end{aligned}
$$

Therefore, from our definition of scaled productivity $a(\jmath) \equiv A(\jmath)^{\frac{1-\nu}{\mu-\nu}}$, aggregate output is

$$Y_t = N_t^{1+\epsilon-\nu}\bar{A}^{1-\nu}\left[K_t^{\alpha}\left(u_tL_t\right)^{1-\alpha}\right]^{\nu} = N_t^{1+\epsilon-\nu}\bar{a}^{\mu-\nu}\left[K_t^{\alpha}\left(u_tL_t\right)^{1-\alpha}\right]^{\nu}. \tag{151}$$

K.14 Aggregate Revenues

Aggregating firm-level revenue shows that it equals to aggregate output. If we combine the optimality condition from the final goods producer problem (i.e. the inverse demand curve) with the aggregate production function, then

$$\begin{align}
\int_0^{N_t} p_t(\imath)y_t(\imath) &= E_t \int_{J_t}^{1} p_t(\jmath)y_t(\jmath)d\jmath \tag{152}\\
&= E_t \int_{J_t}^{1} N_t^{\frac{1+\epsilon-\mu}{\mu}}\left[\frac{y_t(\jmath)}{Y_t}\right]^{\frac{1-\mu}{\mu}} y_t(\jmath)d\jmath \tag{153}\\
&= N_t^{\frac{1+\epsilon-\mu}{\mu}}\left[\frac{1}{Y_t}\right]^{\frac{1-\mu}{\mu}} E_t \int_{J_t}^{1} y_t(\jmath)^{\frac{1}{\mu}}d\jmath \tag{154}\\
&= N_t^{\frac{1+\epsilon}{\mu}}\left[\frac{1}{Y_t}\right]^{\frac{1-\mu}{\mu}} \frac{1}{N}E_t \int_{J_t}^{1} y_t(\jmath)^{\frac{1}{\mu}}d\jmath \tag{155}\\
&= N_t^{\frac{1+\epsilon}{\mu}}\left[\frac{1}{Y_t}\right]^{\frac{1-\mu}{\mu}} \left(Y_tN_t^{-(1+\epsilon)}\right)^{\frac{1}{\mu}} \tag{156}\\
&= Y_t. \tag{157}
\end{align}$$

In the penultimate step we use our definition of aggregate output which is

$$Y_t = N_t^{1+\epsilon}\left[\frac{1}{N_t}\int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}}d\imath\right]^{\mu} = N_t^{1+\epsilon}\left[\frac{1}{N_t}E_t\int_{J_t}^{1} y_t(\imath)^{\frac{1}{\mu}}d\imath\right]^{\mu}.$$

The result implies that at optimality final goods producer profits are zero:

$$\Pi_t^{\mathrm{F}} = 0.$$

And since $\Pi_t^{\mathrm{F}} = \Pi_t$, then

$$\Pi_t = 0.$$

K.15 Aggregate Capital, Labour Demand & Zero-profit Condition

Throughout we use that aggregating firm-level revenues equals to aggregate output as established above.

Aggregate the capital demand condition:

$$r_t k_t(\jmath) = \alpha \frac{\nu}{\mu} p_t(\jmath) y_t(\jmath). \tag{158}$$

Hence, aggregating over all operating firms yields:

$$r_t \int_0^{N_t} k_t(\imath) d\imath = \alpha \frac{\nu}{\mu} \int_0^{N_t} p_t(\imath) y_t(\imath) d\imath \tag{159}$$

$$r_t E_t \int_{J_t}^{1} k_t(\jmath) d\jmath = \alpha \frac{\nu}{\mu} E_t \int_{J_t}^{1} p_t(\jmath) y_t(\jmath) d\jmath \tag{160}$$

$$r_t K_t = \alpha \frac{\nu}{\mu} Y_t \tag{161}$$

$$r_t = \alpha \frac{\nu}{\mu} \frac{Y_t}{K_t}. \tag{162}$$

Aggregate the production-labour demand condition:

$$w_t \ell_t(\jmath) = (1-\alpha) \frac{\nu}{\mu} p_t(\jmath) y_t(\jmath). \tag{163}$$

Hence, aggregating over all operating firms yields:

$$w_t \int_0^{N_t} \ell_t(\imath) d\imath = (1-\alpha) \frac{\nu}{\mu} \int_0^{N_t} p_t(\imath) y_t(\imath) d\imath \tag{164}$$

$$w_t E_t \int_{J_t}^{1} \ell_t(\jmath) d\jmath = (1-\alpha) \frac{\nu}{\mu} E_t \int_{J_t}^{1} p_t(\jmath) y_t(\jmath) d\jmath \tag{165}$$

$$w_t u_t L_t = (1-\alpha) \frac{\nu}{\mu} Y_t \tag{166}$$

$$w_t = (1-\alpha) \frac{\nu}{\mu} \frac{Y_t}{u_t L_t} \tag{167}$$

To aggregate the zero-profit condition, which determines the productivity cut-off

J_t, we first re-express in terms of $\jmath$:

$$\left(1-\frac{\nu}{\mu}\right)p(J_t)y(J_t)=\phi w_t \tag{168}$$

Use the relative productivity result to write

$$\frac{p_t(\jmath)y_t(\jmath)}{p_t(J_t)y_t(J_t)}=\frac{a(\jmath)}{a(J_t)} \quad \therefore \quad p_t(J_t)y_t(J_t)=p_t(\jmath)y_t(\jmath)\frac{a(J_t)}{a(\jmath)}$$

Therefore,

$$\left(1-\frac{\nu}{\mu}\right)p_t(\jmath)y_t(\jmath)a(J_t)=\phi w_t a(\jmath). \tag{169}$$

Hence, aggregating over all operating firms yields:

$$\left(1-\frac{\nu}{\mu}\right)a(J_t)\int_0^{N_t}p_t(\imath)y_t(\imath)d\imath=\phi w_t\int_0^{N_t}a(\imath)d\imath \tag{170}$$

$$\left(1-\frac{\nu}{\mu}\right)a(J_t)E_t\int_{J_t}^{1}p_t(\jmath)y_t(\jmath)d\jmath=\phi w_t E_t\int_{J_t}^{1}a(\jmath)d\jmath \tag{171}$$

$$\left(1-\frac{\nu}{\mu}\right)a(J_t)E_tY_t=\phi w_t E_t(1-J_t)\bar{a}(J_t) \tag{172}$$

$$\left(1-\frac{\nu}{\mu}\right)a(J_t)E_tY_t=\phi w_t N_t\bar{a}(J_t) \tag{173}$$

$$\left(1-\frac{\nu}{\mu}\right)\frac{a(J_t)}{\bar{a}(J_t)}Y_t=w_t(1-u_t)L_t \tag{174}$$

$$\left(1-\frac{\nu}{\mu}\right)\frac{a(J_t)}{\bar{a}(J_t)}\frac{Y_t}{(1-u_t)L_t}=w_t. \tag{175}$$

K.16 Aggregate profits

Aggregate profits are defined net of entry costs. In aggregate they will be zero as expected profits from entering equate to entry costs. There is a distinction between ex-ante profits and ex-post profits. Ex-ante profits are expected profits before entering: these will be zero on average. Whereas, ex-post profits are received after entering and depend on a firm's productivity draw. If the firm has a productivity draw equal to the threshold level of productivity they will make zero profits once entry costs are deducted, if they receive a productivity draw worse than the threshold level (and decide not to produce) they will make negative profits after the entry cost is deducted, and if the firm receives a productivity draw better than the threshold level they will receive positive profits even after the entry cost is deducted.

Aggregate profits are aggregate operating profits less aggregate entry costs

$$\Pi_t = E_t \int_{J_t}^{1} \pi(\jmath) d\jmath - E_t \kappa = 0. \tag{176}$$

It follows that these are zero due to the free entry condition $\int_{J_t}^{1} \pi_t(\jmath) d\jmath = \kappa$. Using goods market clearing, the household budget constraint, labour market clearn and the household budget constraint implies

$$\Pi_t = Y_t - w_t L_t - r_t K_t - E_t \kappa. \tag{177}$$

K.17 General Productivity Distribution Results

If we substitute the aggregate wage equation (30) into the aggregate zero profit condition (31) and use $L_t = 1$, we obtain an expression for u_t in terms of J_t:

$$u_t = \frac{1}{1 + \left(1 - \frac{\nu}{\mu}\right) \frac{\mu}{\nu(1-\alpha)} \frac{a(J_t)}{\bar{a}(J_t)}} \quad 1 - u_t = \frac{\left(1 - \frac{\nu}{\mu}\right) \frac{\mu}{\nu(1-\alpha)} \frac{a(J_t)}{\bar{a}(J_t)}}{1 + \left(1 - \frac{\nu}{\mu}\right) \frac{\mu}{\nu(1-\alpha)} \frac{a(J_t)}{\bar{a}(J_t)}} \tag{178}$$

The term $1 - \frac{\nu}{\mu}$ is the gross profit share in aggregate revenue and $\frac{\mu}{\nu(1-\alpha)}$ is the inverse of labour share in revenue. The ratio of cut-off productivity to average productivity of operating firms $a(J_t)/\bar{a}(J_t)$ determines whether the fraction of labour used in production is increasing or decreasing in the cut-off J_t. This depends on the function that determines the distribution of technology. For the general case, without specifying a productivity distribution, we cannot say whether the ratio is increasing or decreasing in J_t:

$$\frac{a(J_t)}{\bar{a}(J_t)} = \frac{(1 - J_t) a(J_t)}{\int_{J_t}^{1} a(\jmath) d\jmath}. \tag{179}$$

K.18 Pareto Productivity Distribution Results

Consider that the technology variable is Pareto distributed. Given a random variable $\jmath$ drawn from the uniform distribution on the unit interval $(0,1]$, then the technology variable $A(\jmath)$ given by:

$$A(\jmath) = \frac{1}{(1-\jmath)^{\frac{1}{\vartheta}}}. \tag{180}$$

The parameter ϑ is the Pareto shape parameter (tail index). For a given $\jmath \in (0,1]$, a lower shape parameter implies a higher likelihood of a high productivity draw and a lower likelihood of a low productivity draw. We illustrate this graphically in the appendix. Scaled productivity is given by

$$a(\jmath) = A(\jmath)^{\Gamma} = \frac{1}{(1-\jmath)^{\frac{\Gamma}{\vartheta}}}. \tag{181}$$

The term $\Gamma \equiv \frac{1-\nu}{\mu-\nu} \in (0,1)$ is the productivity scaling exponent. If we take the definite integral of scaled productivity over the interval of operating firms, we get

$$\int_{J_t}^{1} a(\jmath) d\jmath = \frac{1}{1-\frac{\Gamma}{\vartheta}}(1-J_t)^{1-\frac{\Gamma}{\vartheta}}. \tag{182}$$

Hence the average productivity of operating firms is

$$\bar{a}(J_t) = \frac{1}{1-J_t}\int_{J_t}^{1} a(\jmath) d\jmath = \frac{1}{1-\frac{\Gamma}{\vartheta}}(1-J_t)^{-\frac{\Gamma}{\vartheta}}. \tag{183}$$

Therefore, average productivity is a linear function of the productivity level of the cut-off firm.

$$\frac{a(J_t)}{\bar{a}(J_t)} = 1 - \frac{\Gamma}{\vartheta}. \tag{184}$$

Average (not-scaled) productivity is

$$\bar{A}(J_t) = \frac{\vartheta}{\vartheta-1}(1-J_t)^{-\frac{1}{\vartheta}}. \tag{185}$$

Average productivity depends on the cut-off and the shape of the Pareto distribution, but not directly on μ or ν (although these will indirectly affect the cut-off J_t).

K.19 Full System

We have specified 26 variables:

$$C_t, K_t, L_t^{\mathrm{s}}, r_t, w_t, \Pi_t, T_t, Y_t, N_t, y_t(\jmath), p_t(\jmath), A(\jmath), a(\jmath), k_t(\jmath), \ell_t(\jmath), \ell_t^{\mathrm{tot}}(\jmath), \pi_t(\jmath), v_t(\jmath),$$
$$J_t, E_t, L_t, u_t, TFP_t, \Pi^{\mathrm{F}}, I_t, \jmath.$$

And there are 26 corresponding equations:

$$L_t^{\mathrm{s}} = 1 \tag{186}$$

$$C_t + I_t = r_t K_t + w_t L_t^{\mathrm{s}} + \Pi_t + T_t \tag{187}$$

$$I_t = K_{t+1} - (1-\delta) K_t \tag{188}$$

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta(r_{t+1} + (1-\delta)) \tag{189}$$

$$Y_t = N_t^{1+\epsilon} \left[\frac{1}{N_t} \int_0^{N_t} y_t(\imath)^{\frac{1}{\mu}} d\imath\right]^{\frac{1}{\mu}} \tag{190}$$

$$p_t(\imath) = N_t^{\frac{1+\epsilon-\mu}{\mu}} \left(\frac{y_t(\imath)}{Y_t}\right)^{\frac{1-\mu}{\mu}} \tag{191}$$

$$y_t(\jmath) = A(\jmath)^{1-\nu} \left[k_t(\jmath)^{\alpha} \ell_t(\jmath)^{1-\alpha}\right]^{\nu} \tag{192}$$

$$\ell_t(\jmath) = \ell_t^{\mathrm{tot}}(\jmath) - \phi \tag{193}$$

$$\frac{w_t}{p_t(\jmath)} = \frac{\nu}{\mu}(1-\alpha)\frac{y_t(\jmath)}{\ell_t(\jmath)} \tag{194}$$

$$\frac{r_t}{p_t(\jmath)} = \frac{\nu}{\mu}\alpha\frac{y_t(\jmath)}{k_t(\jmath)} \tag{195}$$

$$\pi_t(\jmath) = p_t(\jmath) y_t(\jmath) - r_t k_t(\jmath) - w_t \ell_t^{\mathrm{tot}}(\jmath) \tag{196}$$

$$\pi_t(J_t) = 0 \tag{197}$$

$$v_t(\jmath) = \max\{\pi_t(\jmath), 0\} \tag{198}$$

$$\int_0^1 v_t(\jmath) d\jmath = \int_{J_t}^1 \pi_t(\jmath) d\jmath = \kappa \tag{199}$$

$$E_t = \frac{N_t}{1 - J_t} \tag{200}$$

$$K_t = E_t \int_{J_t}^{1} k_t(\jmath) d\jmath \tag{201}$$

$$L_t = E_t \int_{J_t}^{1} \ell_t(\jmath) + \phi \; d\jmath \tag{202}$$

$$u_t = \frac{E_t}{L_t} \int_{J_t}^{1} \ell_t(\jmath) \; d\jmath \tag{203}$$

$$TFP_t = \frac{Y_t}{K_t^{\alpha\nu} L_t^{1+\epsilon-\alpha\nu}} \tag{204}$$

$$Y_t = C_t + I_t \tag{205}$$

$$T_t = E_t \kappa \tag{206}$$

$$\Pi_t = \Pi_t^{\mathrm{F}} \tag{207}$$

$$L_t = L_t^{\mathrm{S}} \tag{208}$$

$$A(\jmath) = \frac{1}{(1 - \jmath)^{\frac{1}{\vartheta}}} \tag{209}$$

$$a(\jmath) = A(\jmath)^{\frac{1-\nu}{\mu-\nu}} \tag{210}$$

$$\jmath \sim Uniform(0,1] \tag{211}$$

K.20 Reduced-form System

In the reduced-form system there is no role for heterogeneity $\jmath$. The distribution of productivity matters through the Pareto shape parameter (ϑ) which affects threshold and average productivity.

$$Y_t - C_t = K_{t+1} - (1-\delta)K_t \tag{212}$$

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta(r_{t+1} + (1-\delta)) \tag{213}$$

$$Y_t = \mathrm{TFP}_t K_t^{\alpha\nu}, \quad \text{where } \mathrm{TFP}_t \equiv \left(\frac{1-u_t}{\phi}\right)^{1+\epsilon-\nu} u_t^{(1-\alpha)\nu} \bar{a}(J_t)^{\mu-\nu} \tag{214}$$

$$r_t = \frac{\nu}{\mu}\alpha\frac{Y_t}{K_t} \tag{215}$$

$$w_t = \frac{\nu}{\mu}(1-\alpha)\frac{Y_t}{u_t} \tag{216}$$

$$\frac{(1-u_t)w_t}{Y_t} = \left(1-\frac{\nu}{\mu}\right)\frac{a(J_t)}{\bar{a}(J_t)} \tag{217}$$

$$\kappa = \phi w_t(1-J_t)\left[\frac{\bar{a}(J_t)}{a(J_t)} - 1\right] \tag{218}$$

$$\frac{a(J_t)}{\bar{a}(J_t)} = 1 - \frac{1}{\vartheta}\left(\frac{1-\nu}{\mu-\nu}\right) \tag{219}$$

$$a(J_t) = \frac{1}{(1-J_t)^{\frac{1}{\vartheta}\left(\frac{1-\nu}{\mu-\nu}\right)}} \tag{220}$$

If we combine the zero-profit condition with the wage equation we get an expression for u_t as a function of the ratio of average productivity to cut-off productivity. Combining this with the Pareto conditon and relative productivity results yields u_t as a function of exogenous parameters. Using this with the aggregate output expression and rental rate expression yields J as a function of K and r.

$$Y_t - C_t = K_{t+1} - (1-\delta)K_t \tag{221}$$

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta(r_{t+1} + (1-\delta)) \tag{222}$$

$$Y_t = \mathrm{TFP}_t K_t^{\alpha\nu}, \quad \text{where } \mathrm{TFP}_t \equiv \left(\frac{1-u_t}{\phi}\right)^{1+\epsilon-\nu} u_t^{(1-\alpha)\nu}\bar{a}(J_t)^{\mu-\nu} \tag{223}$$

$$r_t = \frac{\nu}{\mu}\alpha\frac{Y_t}{K_t} \tag{224}$$

$$u_t = \left[1 + \left(1 - \frac{\nu}{\mu}\right)\frac{\mu}{\nu(1-\alpha)}\frac{a(J_t)}{\bar{a}(J_t)}\right]^{-1} \tag{225}$$

$$\kappa = \phi\frac{\nu}{\mu}(1-\alpha)\frac{Y_t}{u_t}(1-J_t)\left[\frac{\bar{a}(J_t)}{a(J_t)} - 1\right] \tag{226}$$

$$\frac{a(J_t)}{\bar{a}(J_t)} = 1 - \frac{1}{\vartheta}\left(\frac{1-\nu}{\mu-\nu}\right) \tag{227}$$

$$a(J_t) = \frac{1}{(1-J_t)^{\frac{1}{\vartheta}\left(\frac{1-\nu}{\mu-\nu}\right)}} \tag{228}$$

K.21 J as function of K

$$\kappa = \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}(1-J_t)\left(\frac{\bar{a}(J_t)}{a(J_t)} - 1\right)Y_t(J_t, K_t) \tag{229}$$

$$= \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}(1-J_t)\left(\frac{\Gamma}{\vartheta-\Gamma}\right)Y_t(J_t, K_t) \tag{230}$$

$$= \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}(1-J_t)\left(\frac{\Gamma}{\vartheta-\Gamma}\right)TFP_t(J_t)K_t^{\alpha\nu} \tag{231}$$

$$= \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}\left(\frac{\Gamma}{\vartheta-\Gamma}\right)\left(\frac{1-u}{\phi}\right)^{1+\epsilon-\nu} u^{(1-\alpha)\nu}\left(\frac{\vartheta}{\vartheta-\Gamma}\right)^{\mu-\nu}(1-J_t)^{1-\frac{1-\nu}{\vartheta}}K_t^{\alpha\nu} \tag{232}$$

$$= \frac{\phi^{\nu-\epsilon}\nu(1-\alpha)\Gamma\,(1-u)^{1+\epsilon-\nu}\,\vartheta^{\mu-\nu}\,(1-J_t)^{\frac{\vartheta-(1-\nu)}{\vartheta}}\,K_t^{\alpha\nu}}{\mu u^{1-(1-\alpha)\nu}(\vartheta-\Gamma)^{1-\nu+\mu}} \tag{233}$$

Therefore we can rearrange for $J_t(K_t)$:

$$\begin{aligned}
\kappa &= \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}(1-J_t)\left(\frac{\bar{a}(J_t)}{a(J_t)}-1\right)Y_t(J_t,K_t) && (234)\\
&= \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}(1-J_t)\left(\frac{\Gamma}{\vartheta-\Gamma}\right)Y_t(J_t,K_t) && (235)\\
&= \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}(1-J_t)\left(\frac{\Gamma}{\vartheta-\Gamma}\right)TFP_t(J_t)K_t^{\alpha\nu} && (236)\\
&= \phi\frac{\nu}{\mu}(1-\alpha)\frac{1}{u}\left(\frac{\Gamma}{\vartheta-\Gamma}\right)\left(\frac{1-u}{\phi}\right)^{1+\epsilon-\nu}u^{(1-\alpha)\nu}\left(\frac{\vartheta}{\vartheta-\Gamma}\right)^{\mu-\nu}(1-J_t)^{1-\frac{1-\nu}{\vartheta}}K_t^{\alpha\nu} && (237)\\
&= \frac{\phi^{\nu-\epsilon}\nu(1-\alpha)\Gamma\,(1-u)^{1+\epsilon-\nu}\,\vartheta^{\mu-\nu}\,(1-J_t)^{\frac{\vartheta-(1-\nu)}{\vartheta}}K_t^{\alpha\nu}}{\mu u^{1-(1-\alpha)\nu}(\vartheta-\Gamma)^{1-\nu+\mu}} && (238)
\end{aligned}$$

Therefore we can rearrange for $J_t(K_t)$:

$$1-J_t = \left[\frac{\kappa\mu u^{1-(1-\alpha)\nu}(\vartheta-\Gamma)^{1-\nu+\mu}}{\phi^{\nu-\epsilon}\nu(1-\alpha)\Gamma\,(1-u)^{1+\epsilon-\nu}\,\vartheta^{\mu-\nu}K_t^{\alpha\nu}}\right]^{\frac{\vartheta}{\vartheta-(1-\nu)}} \qquad (239)$$

K.22 Aggregate Profits

We can reduce the model conditions to gain expressions for aggregate profits. Aggregating the firm-level profit expression and using the expressions for aggregate capital and labour yields:

$$\int_0^{N_t}\pi_t(\imath)\,d\imath = \int_0^{N_t}p_t(\imath)y_t(\imath)\,d\imath - r_tK_t - w_tL_t. \qquad (240)$$

Using the resource constraint, labour market clearing condition, and household budget constraint yields

$$Y_t = r_tK_t + w_tL_t + \Pi_t + T_t. \qquad (241)$$

Therefore

$$Y_t = \int_0^{N_t}p_t(\imath)y_t(\imath)\,d\imath - \int_0^{N_t}\pi_t(\imath)\,d\imath + \Pi_t + T_t. \qquad (242)$$

Since final output equals to the aggregation of firm-level revenues

$$\Pi_t = \int_0^{N_t} \pi_t(\imath)\, d\imath - T_t. \tag{243}$$

Substituting out the government budget constraint and rewriting the integral yields:

$$\Pi_t = E_t \int_{J_t}^1 \pi_t(\jmath)\, d\jmath - E_t \kappa. \tag{244}$$

We can also obtain the expression for final good producer profits, which is the final goods producer objective function in their optimization problem. Rearranging the household budget constraint and using the profit-clearing condition yields

$$\Pi_t^{\mathrm{F}} = Y_t - r_t K_t - w_t L_t - T_t \tag{245}$$

$$= Y_t - \int_0^{N_t} \pi_t(\imath)\, d\imath - \int_0^{N_t} p_t(\imath) y_t(\imath)\, d\imath - T_t. \tag{246}$$

Using the zero-profit condition $v_t = \int_{J_t}^1 \pi_t(\jmath)\, d\jmath = \kappa$, implies $E_t \int_{J_t}^1 \pi_t(\jmath)\, d\jmath = E_t \kappa$, hence $\int_0^{N_t} \pi_t(\imath)\, d\imath = E_t \kappa$, therefore

$$\Pi_t^{\mathrm{F}} = Y_t - \int_0^{N_t} p_t(\imath) y_t(\imath)\, d\imath. \tag{247}$$

Hence aggregate profits are zero in equilibrium as we have shown that aggregate output equals to the aggregation of firm level revenues.

L Returns to Scale

Returns to scale are a feature of a firm's production function. They describe the change in firm's output as inputs are changed, holding other factors constant.[16] Returns to scale are described as increasing, decreasing or constant depending on whether firm output changes more than, less than or proportionally to a change in inputs. Produc-

[16] Economies of scale are a related concept that capture the cost advantages or disadvantages of production at different scales. Returns to scale, on the other hand, are related to a firm's production function.

tivity measures the amount of output a firm produces for a given amount of factor inputs. At a first-pass, the concepts appear tautological: firms with a production technology that yields more output for a given increase in input should be more productive.

Returns to scale (RTS) are the inverse cost elasticity. The inverse cost elasticity is the ratio of average cost to marginal cost. Thus,

$$\text{RTS} \equiv \left(\frac{\partial \mathcal{C}}{\partial y} \frac{y}{\mathcal{C}} \right)^{-1} = \frac{\mathcal{AC}}{\mathcal{MC}}$$

where AC $\equiv \mathcal{C}/y$ and MC $\equiv \partial \mathcal{C}/\partial y$. We do not observe marginal costs or average costs (total cost divided by output), so directly computing returns to scale using is not possible. Thus we use theory to obtain different expressions for returns to scale. We define returns to scale as increasing, decreasing or constant as follows:

$$\text{RTS} \equiv \begin{cases} \text{Increasing returns,} & \text{if RTS} > 1 \\ \text{Constant returns,} & \text{if RTS} = 1 \\ \text{Decreasing returns,} & \text{if RTS} < 1 \end{cases}$$

Figure 18 presents returns to scale for a firm with U-shaped average cost curve due to upward-sloping marginal cost and a fixed cost, which is consistent with our model. At the intersect of average and marginal cost a firm has constant returns. To the left-hand side of the minimum, average cost exceeds marginal cost so there are increasing returns, and to the right-hand side of the minimum, average cost is less than marginal cost so there are decreasing returns. Hence, size and returns to scale are negatively related at the firm level.

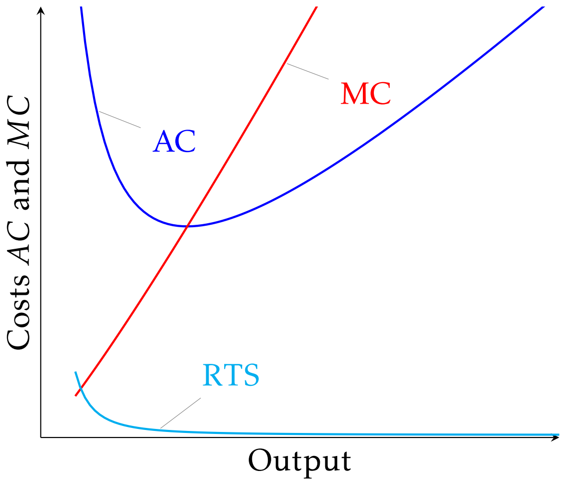


Figure 18: Fixed Cost with Increasing MC, U-Shaped AC Curve

Profit Derivation of RTS

Consider the definition of nominal profits as revenue minus costs

$$\text{Profit} = \text{Price} \times \text{Output} - \text{Cost} = \text{Revenue} - \text{Cost}.$$

From the profit statement, we can see an alternative way of presenting the ratio of average cost to marginal cost which is returns to scale:

$$\frac{\text{AC}}{\text{MC}} = \frac{\text{Price}}{\text{Marginal cost}}\left(1 - \frac{\text{Profit}}{\text{Revenue}}\right).$$

This accounting exercise offers a helpful insight. A firm's returns to scale are its markup multiplied by its profit share remainder (*i.e.* cost share). In the framework we develop, there will be a cut-off firm that makes zero profits. Consequently it has returns to scale equal to its markup and this is the upper bound on returns to scale, as any larger firm will have positive profits and therefore lower returns to scale.

M Comparative Statics: Fixed Cost

How does selection respond to a change in fixed costs? Equation (54) shows that the response of (inverse) selection $1-\tilde{J}$ to ϕ is log-linear:

$$\frac{d\ln(1-\tilde{J})}{d\ln\phi}=\frac{\epsilon-\nu(1-\alpha)}{\vartheta(1-\alpha\nu)-(1-\nu)}\gtreqless 0 \quad\Longleftrightarrow\quad 1+\epsilon\gtreqless 1+\nu(1-\alpha). \tag{248}$$

Given our restrictions on parameters, the denominator is always positive. The numerator is positive or negative depending on the size of external returns to scale $1+\epsilon$. The numerator is negative if $1+\epsilon<1+(1-\alpha)\nu$. Then inverse selection $1-\tilde{J}$ is decreasing in the fixed cost ϕ, so overall the effect on productivity is positive. The intuition is that ϕ increases selection so that only more productive firms surive. This outcomes occurs for two commonly imposed parameter restrictions. First, no external returns to scale $1+\epsilon=1$ and second external returns to scale implicit in a standard CES aggregator $1+\epsilon=\mu$. The first case follows immediately, whereas the second case relies on the parameter restriction $\mu\in(1,1+\nu\min(\alpha,1-\alpha))$ that we impose for profit maximization reasons.

The mechanism for this result is as follows: the free-entry condition is $\kappa=w\phi(\frac{\Gamma}{\Gamma-\vartheta}(1-J))$. As ϕ rises, what happens to the wage? If it perfectly offsets the rise in ϕ, then J is unchanged. But the wage will change if aggregate output $Y=N^{1+\epsilon}\bar{y}$ changes. If ϵ is large enough, a rise in ϕ depresses the wage because it reduces the number of operating firms $N=(1-u)/\phi$. For a large enough ϵ, total overheads $w\phi$ fall, so $1-J$ rises meaning less selection.