Returns to Scale and Productivity in the Macroeconomy

Joel Kariel* Anthony Savagar†

November 15 2022

Find latest draft here.

Abstract

We investigate the puzzle of rising returns to scale but stagnating productivity in several advanced economies. To do this, we develop a model of heterogeneous firms with endogenous returns to scale. We show that growing returns to scale driven by span of control, rather than fixed costs, reduces productivity because it weakens firm selection. Our empirical analysis confirms that productivity and returns to scale are negatively related, and in aggregate returns to scale have increased whilst productivity has weakened. This suggests an important role in recent market structure discussions for technologies that have broadened span of control.

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

Keywords: Returns to Scale, Productivity, Market Structures, Firm Dynamics, Fixed Costs, Span of Control.

*University of Oxford 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 seminar participants at Lancaster, King's, Bath, IFN Stockholm, EEA-ESEM 2022, IAAE 2022, CEF 2022, RES 2022, SNDE 2022, AMEF 2022, SES 2022, MMF 2022, Kent Firm Dynamics Workshop 2021, Exeter Macro Workshop 2022 and Bristol for their helpful comments. Thanks to Omar Licandro, Julian Neira, Tom Schmitz, Petr Sedláček, Danial Lashkari, John Morrow, Anthony Priolo and Kunal Sangani for feedback.

An important factor in recent discussions of rising market power is the growth of returns to scale (Basu 2019; Syverson 2019; De Loecker, Eeckhout, and Unger 2020). Technological improvements, such as cloud computing, enable growing returns to scale. But, whilst scale economies have grown, productivity has stagnated in economies like the US and UK, which are often associated with these technologies. Since scale economies allow firms to increase output more proportionally with inputs and expand at a lower cost, shouldn't this enhance productivity? How can scale economies grow whilst productivity stagnates? In this paper, we develop a theory to understand the relationship between scale economies and productivity. We show that the characteristics of technologies which cause returns to scale also determine the effect on productivity.

Returns to scale can increase through higher fixed costs or lower marginal costs. We explain that the channel matters for aggregate outcomes. Specifically, we show that broadening span of control – which lowers marginal costs – can jointly explain higher returns to scale *and* lower productivity. Whereas, higher fixed costs cause higher returns to scale *and* higher productivity. This occurs because span of control strengthens survival of low-productivity firms, whereas fixed costs weaken survival. Our theory complements applied research which shows that information technologies have widened span of control (Bloom, Garicano, Sadun, and Van Reenen 2014). Our results lead us to conclude that technologies which broaden span of control, such as cloud computing, are important to understand current market structure trends.

For intuition, consider a firm with two employees. If span of control improves, employing a third employee is less costly than it would have been in the past. This helps low productivity firms to survive whilst raising returns to scale.[1] Conversely, if fixed costs rise, returns to scale increase, but productivity increases because low-

[1] In many services-based industries, the cost of an additional employee is similar to the last employee, and significantly less than it would have been in the past. Today a new employee may not need desk-space, physical hard disk memory, machine-specific software licenses, portable USB drives. These innovations are all enabled by cloud-computing. This also helps organizational factors such as staff performance monitoring and staff training, which are closer to Lucas' original interpretation of returns to scale in variable inputs as manager's span of control. This is what we have in mind when we talk about returns to scale in variable production, or span of control.

productivity firms cannot survive. Therefore, two factors that contribute positively to returns to scale have different implications for productivity.

We develop a heterogeneous firm model with endogenous returns to scale, and test the model predictions against estimated returns to scale and productivity for a panel of UK firms. Our theoretical findings show that returns to scale and productivity are negatively related if endogenously higher returns to scale are driven by span of control, whereas there is no relationship between endogenous returns to scale and productivity through changing fixed costs. Our empirical evidence shows a negative relationship between returns to scale and productivity, and we show that overall returns to scale in the UK have increased whilst productivity has declined. Given our theoretical implications, we conclude that this relationship can only be driven by returns to scale based on broadening span of control.

Why does broadening span of control cause productivity to fall? The key mechanism is that span of control improvements make survival easier: wage-denominated fixed costs are easier to pay. Conversely, fixed cost increases make survival harder. This occurs because higher scale economies in variable production reduce expected profits, discouraging entry, which lowers aggregate output and hence wages. Lower wages reduce labour-denominated fixed costs which increases survival of low-productivity firms and 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 dataset is the Annual Business Survey (ABS) which is a representative sample of roughly 60,000 firms in the UK each year 1998-2014. It covers two-thirds of aggregate value-added. A well-known application of the dataset is Aghion, Bloom, Blundell, Griffith, and Howitt (2005).

Related Literature

Several papers link technological improvements to current macroeconomic trends such as rising markups, rising economic profits, declining business dynamism, increasing concentration and declining labour shares (e.g. Bessen 2020; Foster, Haltiwanger, and Tuttle 2022). A seminal paper is Autor, Dorn, Katz, Patterson, and Van Reenen (2017) which assesses the empirical implications of changing technologies for market concentration and the labour share. We provide theory to characterise technologies through returns to scale. And, we find evidence for the vital role of broadening span of control. An important paper for our hypothesis is Bloom, Garicano, Sadun, and Van Reenen (2014). They show that information technology developments have broadened span of control. We have similar innovations in mind when we analyse the theoretical implications of broadening span of control.

Two recent papers develop endogenous growth theory, following Klette and Kortum (2004), to show that changing technology affects firm cost structures, which in turn explains stagnating growth. De Ridder (2019) models intangible inputs, such as software, as reducing marginal costs and raising fixed costs, whilst Aghion, Bergeaud, Boppart, Klenow, and Li (2019) posit a fixed cost that increases with the number of product lines a firm operates. As technology improves, firms become less sensitive to the number of product lines they operate. Our model complements these papers by focusing directly on the implication of changing cost structures for returns to scale. However, we link to productivity through firm selection, following Hopenhayn (1992), rather than an R&D channel. Consistent with their work, we show that falling marginal costs generate lower productivity, and we add how this can occur concurrently with rising returns to scale.

Empirical evidence for France and Spain supports our model implications and empirical evidence for the UK, which shows that returns to scale decrease in firm size. Lashkari, Bauer, and Boussard (2019) study returns to scale with French data, and, like us, they develop a framework with endogenous returns to scale. They present a non-homothetic CES production function with IT and non-IT inputs. Non-homotheticity

in the production function causes firm-level returns to scale to decrease in IT-input cost share. Larger firms have a higher IT share, leading them to have lower returns to scale, whereas smaller firms have higher returns to scale. García-Perea, Lacuesta, and Roldan-Blanco (2021) show that in Spain smaller firms have larger markups driven by higher average costs. Average costs are higher due to higher labour overhead cost shares for smaller firms. This implies smaller firms have higher fixed costs and higher returns to scale.

Our model is a neoclassical growth model with heterogeneous firms based on Hopenhayn and Rogerson (1993), Restuccia and Rogerson (2008), Barseghyan and DiCecio (2011), Barseghyan and DiCecio (2016), Collard and Licandro (2021), and Licandro (2022). We extend this framework to distinguish different sources of endogenous scale economies and we present analytical results for a Pareto distribution. Our formulation of returns to scale is similar to Rotemberg and Woodford (1999), J. Kim (2004), Atkeson and P. J. Kehoe (2005), and D. Kim (2021). These authors recognise the same sources of *internal returns to scale* that we focus on: fixed costs and span of control. They also recognise *external returns to scale* from aggregation which we include for completeness. Recent work by Bilbiie and Melitz (2020), Baqaee, Farhi, and Sangani (2021), and Edmond, Midrigan, and Xu (2021) recognises the importance of external returns to scale – often called love-of-variety in a consumption aggregator – for the welfare implications of firm entry and exit.

Several recent papers provide estimates of returns to scale in the US economy. Gao and Kehrig (2021) estimate returns to scale of 0.96 overall in US manufacturing firms, and returns to scale across 4-digit industries, ranging from 0.86 to 1.3. Ruzic and Ho (2019) use similar US manufacturing data, and 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 for the universe of corporations in France. The most recent estimates of returns to scale in the UK 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.

Lastly, our results are 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).

1 Model

The household side of the model follows a standard neoclassical growth setup. The production side of the economy has firm entry and exit, monopolistic competition, and production functions which have different sources of returns to scale. There are two stages to the firm problem. First, a firm decides whether to pay a fixed, output-denominated, entry cost based on the expected profits they would receive from optimal production decisions. Second, given a firm has entered, it makes optimal production decisions. Upon entering, the firm receives a productivity draw at which point it decides whether to produce or not, and if so how much to produce. The decision to produce or not is based on whether producing output will generate enough revenue to cover a fixed, period-by-period, labour-denominated, overhead cost. At the end of the period all firms die exogenously.

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

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 $1+\epsilon = \mu$, then the aggregator is a standard CES which has 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 condi-

tion 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 as follows. 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 a 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) or span of control. As $\nu \to 1$ the marginal cost curve flattens which raises returns to scale (span of control broadens) and as $\nu \to 0$ the marginal cost curve steepens implying stronger decreasing returns in variable production (span of control weakens). 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

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

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

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:

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

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 which choose to produce 0 after receiving their productivity draw.

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

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

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

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

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

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

We can express aggregate output as:

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

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

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

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 to be active or inactive, measures of entry and active firms $\{E_t, N_t\}_{t=0}^{\infty}$; 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 inactive, and demand factors according to (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 present the steady state.

2.1 Productivity with Pareto Distribution

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

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

where ϑ is the Pareto tail parameter, and we define the productivity scaling exponent as

$$\Gamma \equiv \frac{1-\nu}{\mu-\nu} \in (0,1). \tag{37}$$

Average productivity is a linear function of the productivity level of the cut-off firm:

$$\frac{a(J_t)}{\bar{a}(J_t)} = 1 - \frac{\Gamma}{\vartheta}. \tag{38}$$

We can express average productivity as:

$$\bar{A}(J_t) = \frac{\vartheta}{\vartheta-1}(1-J_t)^{-\frac{1}{\vartheta}} \tag{39}$$

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

2.2 Returns to Scale

Heterogeneous productivity leads to heterogeneous firm output, and in turn heterogeneous returns to scale because the fixed cost is the same for all 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 dominate the increasing returns from the overhead, so overall the firm has decreasing returns.

2.2.1 External Returns to Scale

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

From equations (9) and (10), the responses of firm output to a change in each variable input are as follows:

$$\frac{\partial \ln y_t(j)}{\partial \ln k_t(j)} = \nu\alpha, \quad \frac{\partial \ln y_t(j)}{\partial \ln \ell_t(j)} = \nu(1-\alpha).$$

Therefore, changing capital or changing production labour have constant effects on output regardless of firm size. Returns to scale in variable production measures the response of firm output to variable inputs:

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

The effect of a change in total labour input, which is what we can measure in the data, is decreasing in firm size:

$$\frac{\partial \ln y_t(\jmath)}{\partial \ln \ell_t^{\text{tot}}(\jmath)} = \nu(1-\alpha)\left(1+\frac{\phi}{\ell_t(\jmath)}\right) \quad \in (\nu(1-\alpha), \mu - \alpha\nu).$$

The lower bound is reached by the largest firm $\ell_t(\jmath) \to \infty$ and the upper bound is reached by the zero-profit firm which has the largest overhead to production labour ratio, given by

$$\frac{\phi}{\ell_t(J_t)} = \frac{1}{1-\alpha}\left(\frac{\mu}{\nu} - 1\right).$$

In sum, the response of firm output to total labour is smaller for larger firms. To get an intuition for this result, consider a firm that has 10 total labour units, where 9 go to overheads and 1 goes to production, total labour increases by 10% to 11, overhead labour remains at 9 therefore production labour increases 100% to 2, hence output responds to the 100% increase in production labour rather than the 10% increase in total labour. Thus, at a small scale an increase in total labour has a big effect on output. Conversely, consider a firm that has 9 units of overhead labour and 991 units of production labour, a 10% increase in total labour from 1,000 to 1,100 also raises production labour by 10% from 991 to 1,091 and correspondingly output responds the same as if production labour were directly increased.

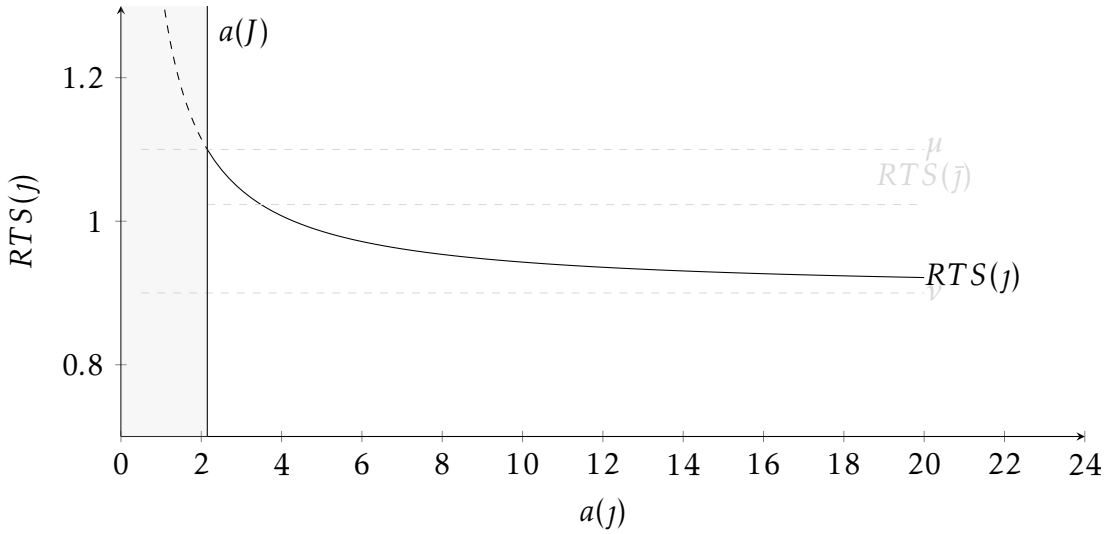
Overall returns to scale at the firm-level – which is what we can measure in the data – is the response of firm output to a change in all inputs:

$$\text{RTS}_t(\jmath) \equiv \frac{\partial \ln y_t(\jmath)}{\partial \ln k_t(\jmath)} + \frac{\partial \ln y_t(\jmath)}{\partial \ln \ell_t^{tot}(\jmath)} = \nu\left(1+(1-\alpha)\frac{\phi}{\ell_t(\jmath)}\right) = \nu + (\mu-\nu)\frac{a(J_t)}{a(\jmath)}. \tag{40}$$

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. 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 $a(\jmath)$. 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 share in production labour becomes negligible.

Figure 1: Firm-level Returns to Scale



Plot shows returns to scale of a firm given its productivity draw. In the shaded region firms are inactive and the dashed line shows their hypothetical returns to scale if they were to produce. The horizontal lines show the bounds on returns to scale of active firms $RTS(j) \in (\nu, \mu)$ and the average returns to scale of active firms given in equation (41). The $RTS(j)$ plot corresponds to equation (40) where parameter values are $\phi = 0.5, \nu = 0.9, \mu = 1.1, \vartheta = 1.3$, and we specify the endogenous cut-off to $J = 0.1$ which is the steady-state equilibrium value of the cut-off that we solve for numerically in our main exercise.

2.2.3 Average Returns to Scale

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

$$RTS(\bar{j}) = \mu - \frac{1-\nu}{\vartheta}. \tag{41}$$

Returns to scale of the average firm tends towards the markup (which is the returns to scale of the smallest, zero-profit, firm) as ϑ gets large. This is because a higher Pareto shape parameter causes a higher density of firms in the left-hand side of the productivity distribution and a thinner right-hand tail. This implies a higher share of low-productivity firms which have high RTS, close to the markup.

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:

$$u = \left[1 + \frac{1}{\nu(1-\alpha)}\left(\mu - \nu - \frac{1-\nu}{\vartheta}\right)\right]^{-1}. \tag{42}$$

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

$$TFP_t(J_t) = \left(\frac{1-u}{\phi}\right)^{1+\epsilon-\nu} u^{(1-\alpha)\nu} \bar{a}(J_t)^{\mu-\nu} \tag{43}$$

$$= \left(\frac{1-u}{\phi}\right)^{1+\epsilon-\nu} u^{(1-\alpha)\nu} \left(\frac{\vartheta}{\vartheta-\Gamma}\right)^{\mu-\nu} (1-J_t)^{-\frac{1-\nu}{\vartheta}}. \tag{44}$$

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

$$Y_t(J_t, K_t) = TFP_t(J_t) K_t^{\alpha\nu}. \tag{45}$$

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:

$$1 - J_t = \left[\frac{\kappa\mu u^{1-(1-\alpha)\nu}(\vartheta-\Gamma)^{1-\nu+\mu}}{\phi^{\nu-\epsilon}\nu(1-\alpha)\Gamma(1-u)^{1+\epsilon-\nu}\vartheta^{\mu-\nu}K_t^{\alpha\nu}}\right]^{\frac{\vartheta}{\vartheta-(1-\nu)}} \tag{46}$$

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:

$$r_t = \alpha\frac{\nu}{\mu}\frac{Y_t(K_t)}{K_t}. \tag{47}$$

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

$$Y_t(K_t) - C_t = K_{t+1} - (1-\delta)K_t \tag{48}$$

$$\left(\frac{C_{t+1}}{C_t}\right)^{\sigma} = \beta(r_{t+1}(K_{t+1}) + (1-\delta)). \tag{49}$$

2.4 Steady state

A steady-state equilibrium is an equilibrium in which prices, quantities, entry and firm productivity are constant. Tilde notation represents a variable in steady state. In steady state, denoted with a tilde, the consumption Euler equation implies that

$$\tilde{r} = \frac{1}{\beta} + (1-\delta). \tag{50}$$

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:

$$1 - \tilde{J} = \left[\mu\left(\frac{\kappa}{1-\nu}\right)^{1-\alpha\nu}\left(\frac{\tilde{r}}{\alpha\nu}\right)^{\alpha\nu}\left[\frac{\mu-\nu}{\nu(1-\alpha)\vartheta}\right]^{(1-\alpha)\nu}(\vartheta-\Gamma)^{\mu-\alpha\nu}\phi^{\epsilon-(1-\alpha)\nu}(1-u)^{-\epsilon}\vartheta^{1-\mu+(1-\alpha)\nu}\right]^{\frac{\vartheta}{\vartheta(1-\alpha\nu)-(1-\nu)}} \tag{51}$$

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

3 Theoretical Results

We analyse the steady-state of the model and discuss the implications of changing fixed costs ϕ and span of control ν for returns to scale and productivity. We show that only rising span of control ν can jointly increase returns to scale and decrease productivity.

3.1 Comparative Statics: Average Returns to Scale

In our empirical analysis we measure average returns to scale of firms within an industry, rather than the returns to scale for each firm, hence it is the relevant object for our comparative statics analysis. In the data we only observe firms who produce, which in the theory means we observe firms with $\jmath \in (J_t, 1)$.We pool together all firm-time observations by industry. We estimate production functions for each industry which yields a single set of regression coefficients for each industry. These coefficients represent the returns to scale of the average firm in the industry.

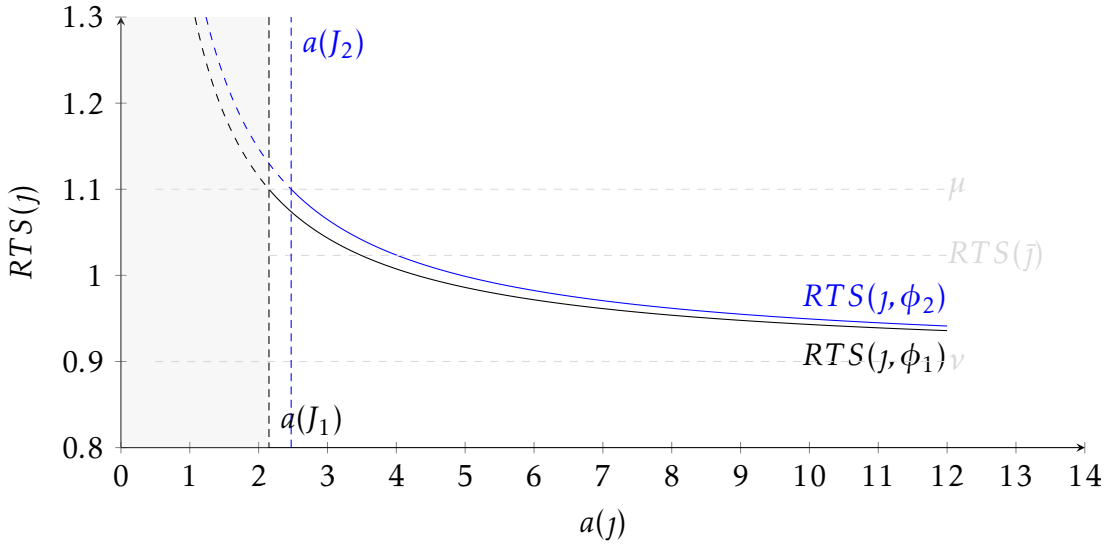
3.1.1 The Effect of Fixed Costs on Returns to Scale

From equation (41), 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, $RTS(\jmath)$ in equation (40), on average returns to scale do not change. This occurs because J_t adjusts so the extensive margin effect (selection of firms) cancels-out the intensive margin effect (effect on individual firms' RTS). For example, a higher fixed cost raises selection J_t such that surviving firms have a higher level of production labour, hence the overhead cost to production labour ratio in (40) remains unchanged:

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

Figure 2 shows the impact of a rise in the fixed cost ϕ on selection, firm-level returns to scale and average returns to scale. Higher ϕ raises returns to scale for all incumbents (the $RTS(\jmath)$ curve shifts up), but increased selection (a higher cut-off J) eliminates the least-productive firms, with the highest returns to scale. Hence average returns to scale RTS($\bar{\jmath}$) is unchanged. The change in fixed cost does not affect the upper and lower bound on returns to scale which remains $RTS(\jmath) \in (\nu, \mu)$.

Figure 2: Firm-level Returns to Scale, Change in ϕ



Plot shows returns to scale of a firm given its productivity draw from a Pareto distribution. In the shaded region firms are inactive and the dashed line shows their hypothetical returns to scale if they were to produce. The horizontal lines show the bounds on returns to scale of active firms $RTS(\jmath) \in (\nu, \mu)$ and the average returns to scale of active firms given in equation (41). The $RTS(\jmath)$ plot corresponds to equation (40) where parameter values are $\nu = 0.9$, $\mu = 1.1$, $\vartheta = 1.3$. We solve for the endogenous productivity cut-offs $a(J_1), a(J_2)$ when $\phi_1 = 0.5, \phi_2 = 0.75$.

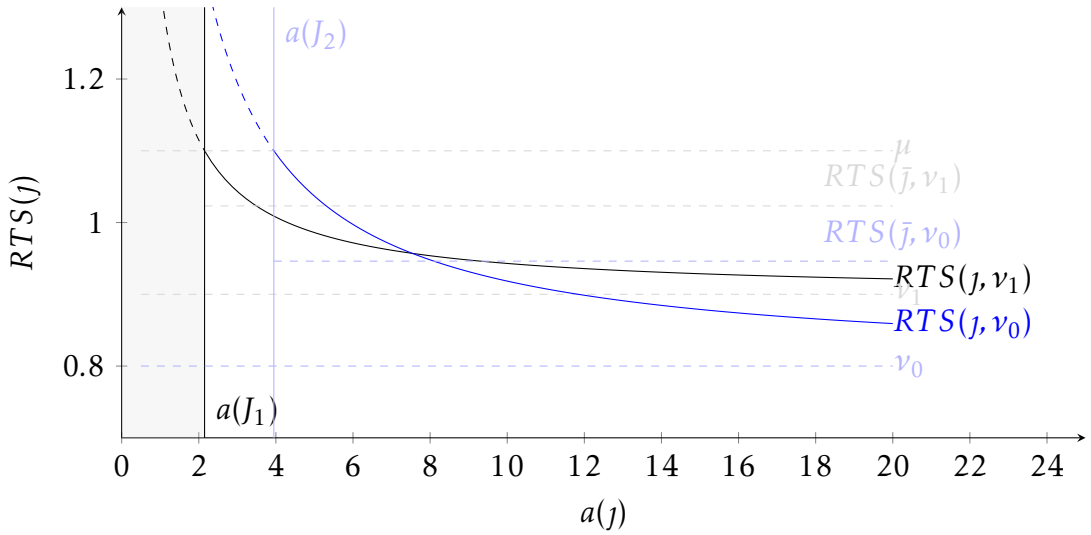
3.1.2 The Effect of Span of Control on Returns to Scale

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

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

A rise in span of control ν increases average returns to scale unambiguously. However, the effect on firm-level returns to scale depends on the productivity draw. Figure 3 shows that the returns to scale schedules intersect. To the left-hand side of the intersect, the rise in ν reduces returns to scale and to the right-hand side of the intersect the rise in ν increases returns to scale. The upper bound on returns to scale is fixed but the lower bound increases. Therefore the range of returns to scale declines.

Figure 3: Firm-level Returns to Scale, Change in ν



Plot shows returns to scale of a firm given its productivity draw. In the shaded region firms are inactive and the dashed line shows their hypothetical returns to scale if they were to produce. The horizontal lines show the bounds on returns to scale of active firms $RTS(\jmath) \in (\nu, \mu)$ and the average returns to scale of active firms given in equation (41). The $RTS(\jmath)$ plot corresponds to equation (40) where parameter values are $\phi = 0.5$, $\mu = 1.1$, $\vartheta = 1.3$. We solve for the endogenous productivity cut-offs $a(J_1), a(J_2)$ when $\nu_1 = 0.9, \nu_0 = 0.8$.

3.2 Comparative Statics: Productivity

Given equation (39), changes in exogenous parameters such as ν and ϕ, denoted x below, affect average productivity through the degree of selection:

$$\frac{d \ln \bar{A}(\tilde{J})}{d \ln x} = \frac{\partial \ln \bar{A}}{\partial \ln(1-\tilde{J})} \times \frac{d \ln(1-\tilde{J})}{d \ln x} = -\frac{1}{\vartheta} \times \frac{d \ln(1-\tilde{J})}{d \ln x}. \tag{52}$$

The first term captures that greater selection ($1 - \tilde{J} \to 0$) always increases average productivity. The second term captures how selection (51) responds to a change in an exogenous parameter.

3.2.1 The Effect of Fixed Costs on Average Productivity

Equation (51) shows that the response of (inverse) selection $1 - \tilde{J}$ to ϕ is log-linear:

$$\frac{d \ln (1-\tilde{J})}{d \ln \phi} = \frac{\epsilon - \nu(1-\alpha)}{\vartheta(1-\alpha\nu) - (1-\nu)} \gtreqless 0 \quad \iff \quad 1 + \epsilon \gtreqless 1 + \nu(1-\alpha). \tag{53}$$

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 survive. This outcome 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.

A rise in fixed cost ϕ can reduce selection if it decreases labour overhead payments. The free-entry condition $\kappa = \tilde{w}\phi(\frac{\Gamma}{\Gamma-\vartheta}(1-\tilde{J}))$, where steady-state wage depends on ϕ, shows that if steady-state overhead payments, $\tilde{w}\phi$, decrease then $\tilde{J}$ must fall to maintain equality. A rise in ϕ decreases wages if it reduces aggregate output sufficiently. A fall in aggregate output occurs if expected profits decline, reducing entry and therefore the number of operating firms, in reduced-form $N = (1-u)/\phi$. A fall in the number of operating firms will have a large negative effect on aggregate output if $1+\epsilon$ is large because $Y = N^{1+\epsilon}y(\bar{j})$.

3.2.2 The Effect of Span of Control on Average Productivity

The steady-state expression of (inverse) selection $1-\tilde{J}$ is nonlinear in ν. We cannot provide a closed-form result. Figure 4 shows the steady-state productivity cut-off $\tilde{J}$ for values of ν from 0.75 to 0.90. Exogenous parameters are set as in Table 1.

Table 1: Parameter Values for Comparative Statics

	Parameter	Value	Target
μ	Markup estimate	1.1	ONS (2022)
α	Capital share	0.25	ABS (authors' calculations)
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 ONS (2022). The value of α implies a capital share of 0.25ν, to obtain values close to our calculations in the ABS. The value of κ follows Barseghyan and DiCecio (2011). External returns to scale 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 $\tilde{r} = \frac{1}{\beta} + 1 - \delta$.[3] The Pareto shape ϑ is calibrated to match the firm size distribution, as in Table 14 in the Appendix. The overhead cost ϕ is set such that the proportion of active and inactive firms ($\tilde{J}$) is empirically plausible.

[3]Data on UK long-term government bond and inflation used to compute the real interest rate from FRED database: IRLTLT01GBM156N and FPCPITOTLZGGBR.

Figure 4: Cut-off Productivity $\tilde{J}$ and Variable Returns to Scale ν

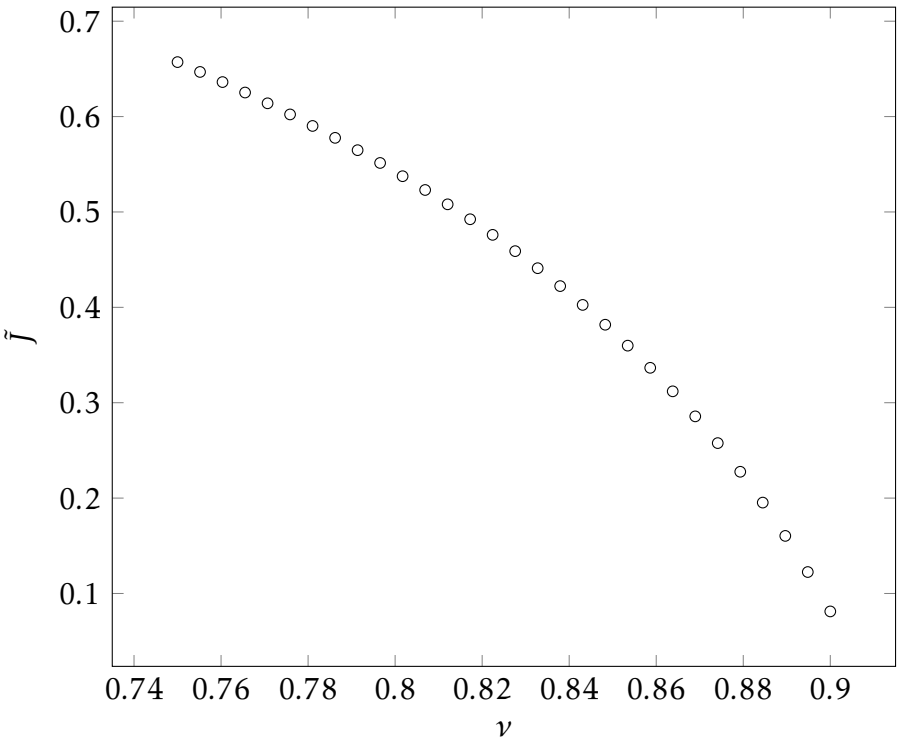


Figure 5: Average Productivity $A(\tilde{J})$ and Variable Returns to Scale ν

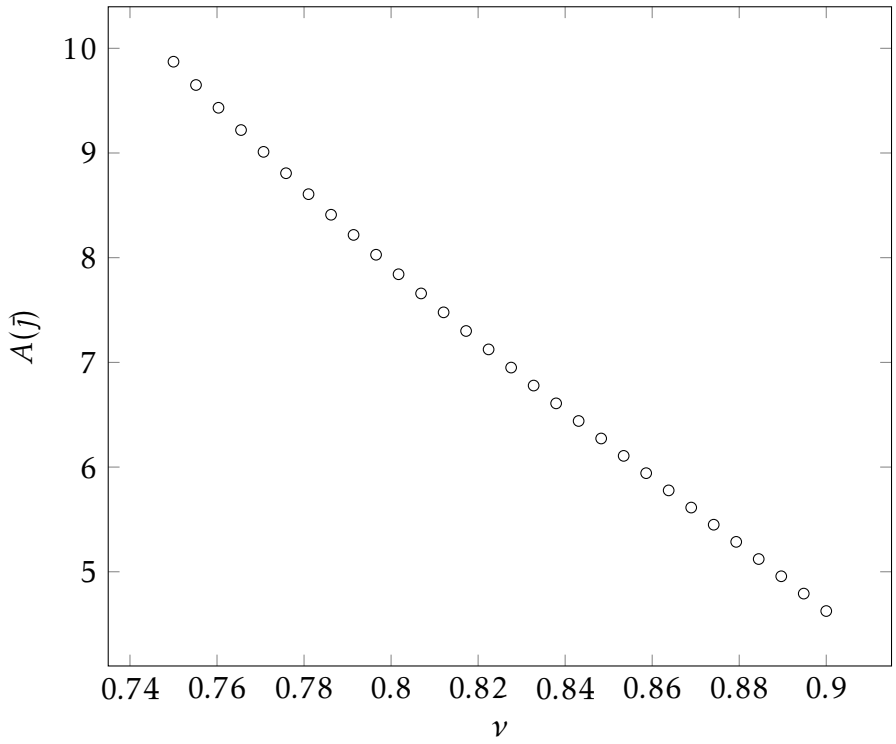
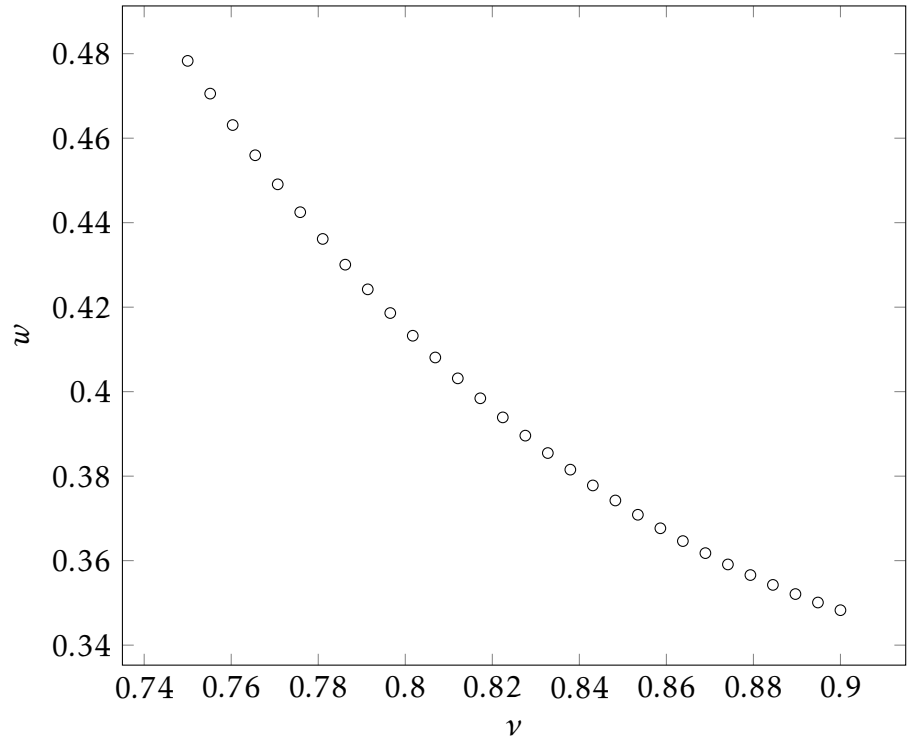


Figure 4 shows that the cut-off productivity is declining in variable returns to scale

ν, consequently by (52) average productivity is falling (Figure 5). This occurs because the wage is declining in ν (Figure 6), such that payments to overheads $w_t\phi$ are falling in ν, therefore less productive firms can pay the fixed cost and survive with higher ν.

Figure 6: Scatter plot of steady-state wage and variable returns to scale ν



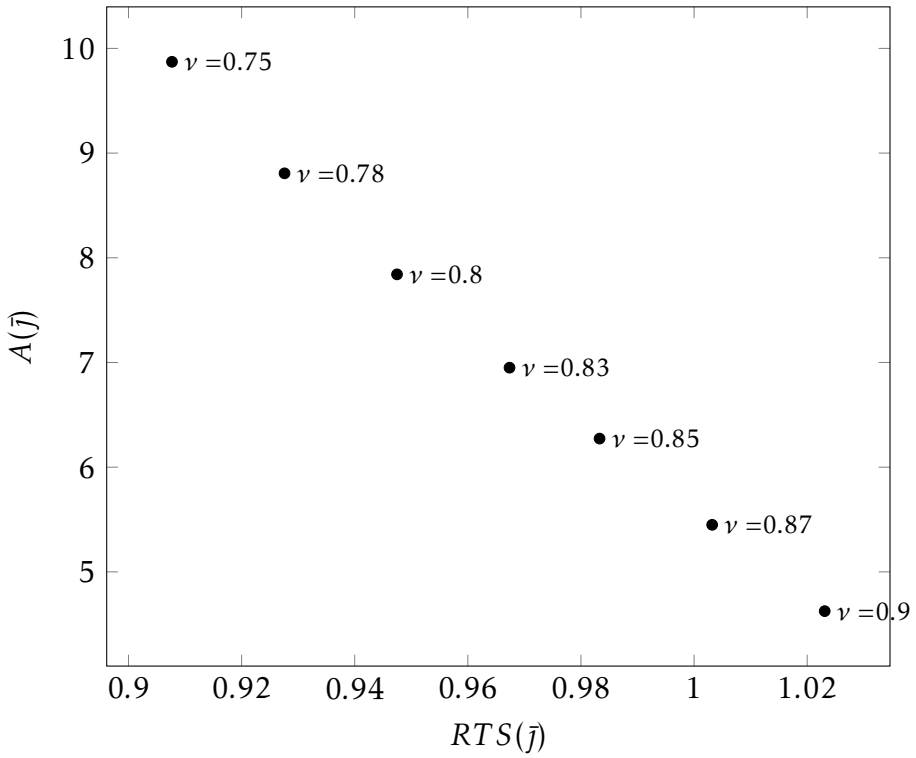
The negative relationship between the w and ν occurs because higher ν reduces the number of active firms which in turn lowers aggregate output and consequently wage. We present figures that show the negative relationship between ν and steady-state N and E in the appendix. The number of operating firms is a function of exogenous parameters $N = (1-u)/\phi$, given u in equation (42). A rise in ν decreases utilization of overhead labour $1-u$. Hence the number of operating firms falls in ν.

3.3 Comparative Statics: Returns to Scale and Productivity

Figure 7 shows that a rise in ν lowers average productivity, but raises average returns to scale. As ν rises, the returns to scale of the average firm increases, given Equation (41). However there is *less* selection with higher ν, so average productivity falls,

as in Figure 5. This gives a negative relationship between returns to scale and productivity as ν increases.

Figure 7: Labelled scatter plot of average productivity and average returns to scale



3.4 Theoretical Results Summary

Our analysis shows how fixed costs and span of control affect measured returns to scale. Changes in these two parameters have different implications for average returns to scale and productivity. Fixed costs do not affect average returns to scale, due to selection counter-acting individual firm effects, whereas span of control positively affects average returns to scale. In general, both channels have ambiguous effects on productivity, but under common restrictions ϕ is positively related to productivity due to stronger selection, whilst, in numerical simulations, ν is negatively related due to weaker selection. Consequently, technologies characterised by increasing span of control can jointly explain higher returns to scale, but lower productivity.

3.4.1 Additional Model Features

The model has several salient features. Returns to scale are decreasing in firm size. The smallest firm has returns to scale equal to the markup, whilst the largest firm has returns to scale equal to the span of control. Therefore, changes in span of control affect the range of returns to scale, whereas changes in fixed costs do not. All firms charge the same constant markup, hence differences in size, and consequently returns to scale, are reflected in a firm's economic profit share in revenue. The smallest firm, which has highest returns to scale, has zero profits whilst the largest firm, which has lowest returns to scale, has large positive profits as a fraction of revenue. Appendix D provides empirical support of the negative relationship between firm size and returns to scale.

4 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.

4.1 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 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. [4]

4.2 Mapping Production Function to 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:

$$\ln p_t(\jmath)y_t(\jmath) = \beta_1 \ln k_t(\jmath) + \beta_2 \ln \ell_t^{\text{tot}}(\jmath) + \varepsilon_t(\jmath). \tag{54}$$

4.2.1 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 $\jmath = 1 \ldots N$ within each industry and common across time. The coeffi-

[4] 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 as possible with our theoretical model.

cients are

$$\beta_1 = \frac{\partial \ln p_t(\jmath) y_t(\jmath)}{\partial \ln k_t(\jmath)} = \frac{\nu}{\mu}\alpha \tag{55}$$

$$\beta_2 = \frac{\partial \ln p_t(\jmath) y_t(\jmath)}{\partial \ln \ell_t^{\text{tot}}(\jmath)} = \frac{\nu}{\mu}(1-\alpha)\left(1 + \frac{\phi}{\ell(\bar{\jmath})}\right) = \frac{1}{\mu}\left(\mu - \nu\alpha - \frac{1-\nu}{\vartheta}\right). \tag{56}$$

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 (56). 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:

$$\beta_1 + \beta_2 = \frac{\nu}{\mu}\left(1 + (1-\alpha)\frac{\phi}{\ell(\bar{\jmath})}\right) = 1 - \frac{1-\nu}{\mu\vartheta}. \tag{57}$$

4.2.2 Estimated Technical Efficiency

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

$$\ln \hat{TFPR}_t(\jmath) = \ln p_t(\jmath) y_t(\jmath) - \hat{\beta}_1 \ln k_t(\jmath) - \hat{\beta}_2 \ln \ell_t^{\text{tot}}(\jmath). \tag{58}$$

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 $\ln \hat{TFPR}_t(\jmath)$ is the estimated revenue function residual for a specific firm given their capital, labour and revenue, teamed with industry average revenue elasticity coefficients. Our $\ln \hat{TFPR}_t(\jmath)$ 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 $\ln \bar{TFPR}(\imath)$ 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

$$\ln \bar{TFPR} = \frac{\nu}{\mu}(1-\alpha)\left[\ln \ell_t(\bar{\jmath}) - \frac{\ell_t^{\text{tot}}(\bar{\jmath})}{\ell_t(\bar{\jmath})}\ln \ell_t^{\text{tot}}(\bar{\jmath})\right] + \frac{1-\nu}{\mu}\ln A(\bar{\jmath}) + \frac{\mu-1}{\mu}\ln Y_t + \frac{1+\epsilon-\mu}{\mu}\ln N_t. \tag{59}$$

By studying differences in $\bar{TFPR}$ 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.[5] In Appendix C, we show that $\ln \bar{TFPR}$ in our model is declining in ν, and that this is mostly driven by falling true average productivity $\bar{A}$. Therefore, $\ln \bar{TFPR}$ is a good proxy for $\ln \bar{A}$.

4.2.3 Relationship between Estimated RTS and TFPR

We want to understand the relationship between the average $\ln \bar{TFPR}$ of firms in an industry and the sum of the estimated coefficients for each industry, which represent returns to scale. To do this we run the regression:

$$\ln \bar{TFPR}(\imath) = \gamma_0 + \gamma_1\left[\hat{\beta}_1(\imath) + \hat{\beta}_2(\imath)\right] \tag{60}$$

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

5 Empirical Results

Our model makes two predictions. First, average returns to scale of operating firms are invariant to fixed costs, but positively related to variable costs. Second, overall returns

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

to scale are negatively related to productivity. Our results confirm these predictions.

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. We will refer to our estimates as 'returns to scale', although technically they are revenue elasticities. Given our model assumes CES demand and monopolistic competition, markups are constant across firms and over time, so revenue elasticity variation is proportional to returns to scale variation. Our main findings are:

1. On average returns to scale in the UK are 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.

5.1 Returns to Scale Levels

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 UK from 1998 - 2014 are 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 produce below their *minimum efficient scale*, as average costs exceed marginal costs.

We favour the ACF and GNR results as the former identifies labour elasticities, and the latter sidesteps this issue with use of the first-order condition on materials. ACF matches more closely to the Cobb-Dogulas value-added production function in our theory so this is our primary methodology. 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.

5.1.1 Additional Estimates by 2-Digit Sector and by Firm Size

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.

Finally, we split the sample of firms into (approximately) the bottom and top quintiles by employment. Equation (40) highlights that returns to scale must be bounded between variable returns to scale ν and the markup μ. Estimating returns to scale on these two groups gives an approximation of these two parameters.[6] We find that the largest employment quintile of firms has returns to scale equal to 1.31, while the smallest employment quintile has returns to scale of 0.97. Table 13 includes further information on the number of firms, average employment statistics, and estimated $\ln TFPR$.

5.2 Returns to Scale Growth

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 fol-

[6] We estimate returns to scale on a large enough set of firms such that we get precise estimates. However, naturally we include firms that are not large/small enough to actually have returns to scale at the boundary conditions. Therefore our estimates will be smaller/larger than the true boundary conditions.

lowing ACF. Underlying coefficients are found in Table 11. Estimates with GNR are provided in the Appendix Table 12. Economy-wide, there is 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 have also increased across other macroeconomic sectors. 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.

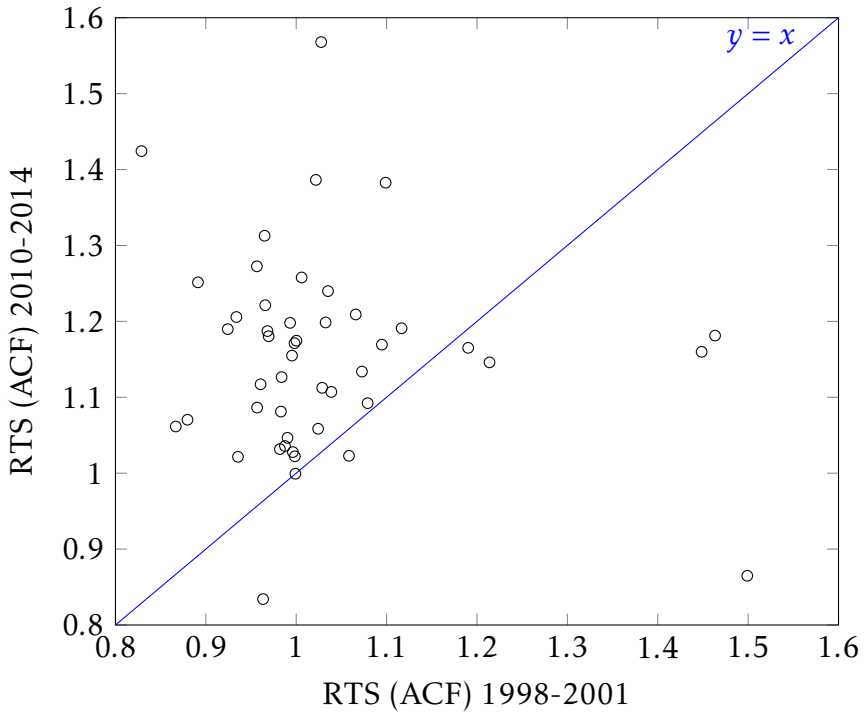
	1998 - 2001	2002 - 2005	2006 - 2009	2010 - 2014
		Economy-Wide		
RTS	0.988	1.081	1.046	1.061
N	153,874	144,465	108,619	120,855
		Manufacturing		
RTS	1.11	1.24	1.10	1.15
N	41,572	36,074	24,280	21,626
		Construction		
RTS	0.91	0.87	1.08	1.29
N	13,050	13,180	9,797	14,145
		Wholesale/Trade/Transport		
RTS	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 8 plots a comparison of returns to scale in 1998 - 2001,

compared to 2010 - 2014, across sectors, using the ACF estimation.[7] GNR estimates are provided in Figure 16 in the Appendix. Most industries experienced an increase in returns to scale, as the majority of points sit above the 45 degree line.

Figure 8: Changing Returns to Scale by 2-digit SIC, ACF Estimation.



Comparison of returns to scale at 2-digit SIC level, from 1998 - 2001 to 2010 - 2014. Line is 45 degree line: points above that line are consistent with a rise in returns to scale.

5.3 Returns to Scale & Productivity

We estimate TFPR using control function methods. Figure 9 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 over from 2012 to 2014. This result holds across estimation methods. See Figure 15 in the Appendix for Gandhi, Navarro, and Rivers (2020).

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

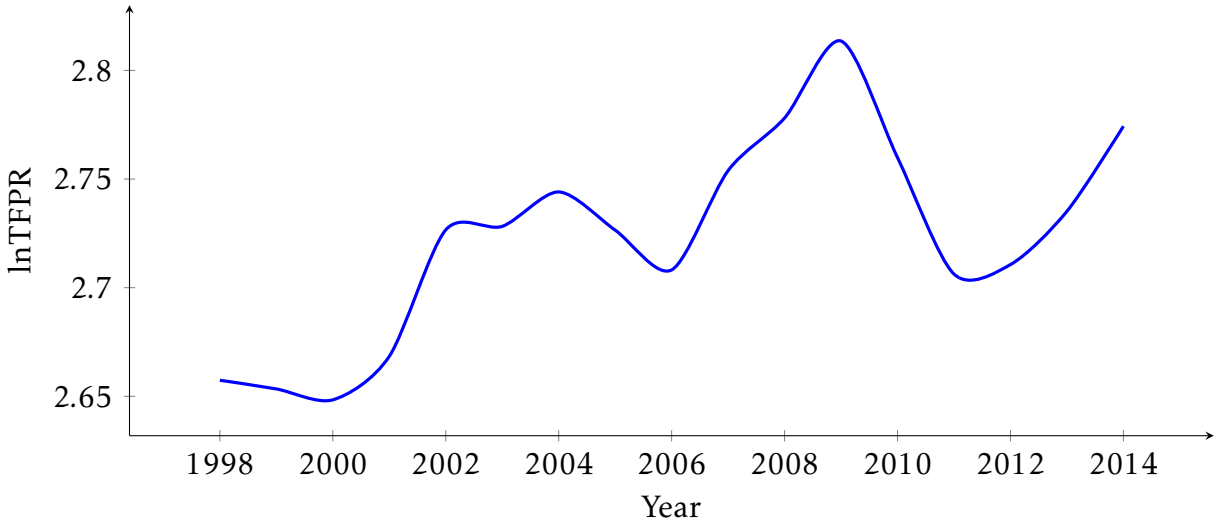


Figure 9: Aggregate TFPR (ACF)

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 10 contains a scatter plot of returns to scale and average log TFPR across 2-digit sectors for ACF estimation. Each point represents the average firm within the sector. 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. We obtain the same general result with GNR estimation, plotted in Figure 17 in the Appendix.

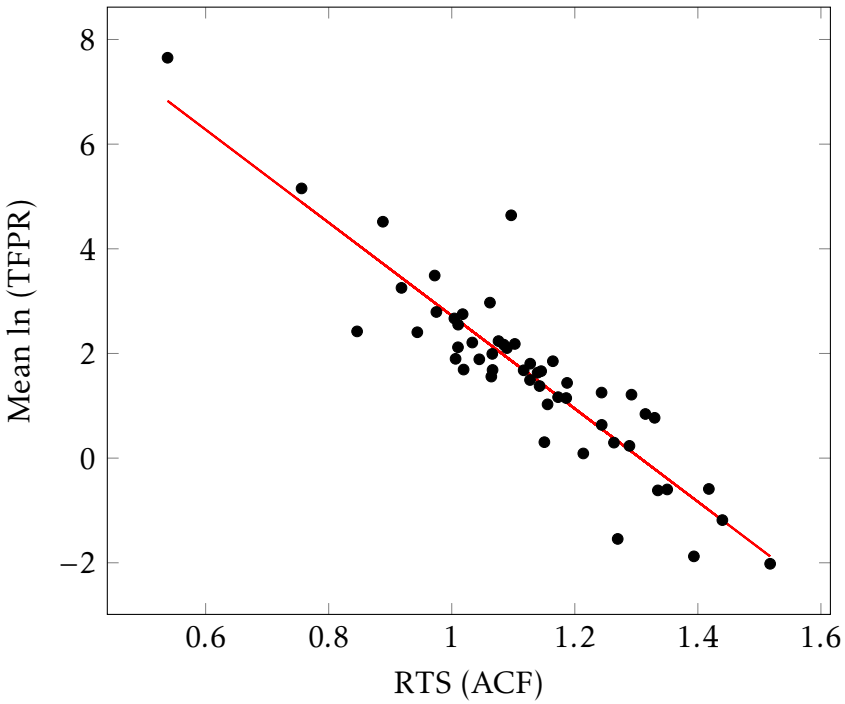


Figure 10: Relationship between Revenue Elasticity and TFP, ACF

Table 4 contains estimates from a regression of returns to scale on average log TFPR across 2-digit sectors. 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.

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.*

5.3.1 Robustness to TFPR Proxy

According to Section 4.2, 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. To do this, we pool the observations from Table 4, dropping sectors that do not appear in all four time periods. This yields the number of observations in Table 5. The first column presents the estimated coefficient from a 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. 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.*

5.4 Empirical Results Summary

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 have 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.

6 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 ('span of control') 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 rising 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, Nicholas 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.

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

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.

— (2016). "Externalities, endogenous productivity, and poverty traps". In: *European Economic Review* 85.C, pp. 112–126.

Basu, Susanto (2019). "Are price-cost markups rising in the United States? A discussion of the evidence". In: *Journal of Economic Perspectives* 33.3, pp. 3–22.

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

Bilbiie, Florin O. and Marc J Melitz (Dec. 2020). *Aggregate-Demand Amplification of Supply Disruptions: The Entry-Exit Multiplier*. Working Paper 28258. National Bureau of Economic Research.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen (2014). "The distinct effects of information technology and communication technology on firm organization". In: *Management Science* 60.12, pp. 2859–2885.

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 Gabriel Unger (Jan. 2020). "The Rise of Market Power and the Macroeconomic Implications*". In: *The Quarterly Journal of Economics* 135.2, pp. 561–644.

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.

Foster, Lucia S, John 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)*.

García-Perea, Pilar, Aitor Lacuesta, and Pau Roldan-Blanco (2021). "Markups and cost structure: Small Spanish firms during the Great Recession". In: *Journal of Economic Behavior & Organization* 192, pp. 137–158.

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.

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 (1992). "Entry, Exit, and Firm Dynamics in Long Run Equilibrium". In: *Econometrica* 60.5, pp. 1127–1150.

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.

Kim, Jinill (2004). "What determines aggregate returns to scale?" In: *Journal of Economic Dynamics and Control* 28.8, pp. 1577–1594.

Klette, Tor Jakob and Samuel Kortum (2004). "Innovating firms and aggregate innovation". In: *Journal of Political Economy* 112.5, pp. 986–1018.

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.

ONS (2022). *Estimates of markups, market power and business dynamism from the Annual Business Survey, Great Britain: 1997 to 2019*. Tech. rep. Office for National Statistics website.

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.

Rotemberg, Julio J. and Michael Woodford (1999). "The cyclical behavior of prices and costs". In: *Handbook of Macroeconomics*. Ed. by J. B. Taylor and M. Woodford. Vol. 1. Handbook of Macroeconomics. Elsevier. Chap. 16, pp. 1051–1135.

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

Syverson, Chad (2019). "Macroeconomics and market power: Context, implications, and open questions". In: *Journal of Economic Perspectives* 33.3, pp. 23–43.

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.[8] However, for us the most important documentation regards the ABS which is a core ONS product used in the construction of national accounts, and consequently well-documented. 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

[8] Specifically, the ARD is a dataset constructed for researchers, rather than for national accounts purposes. It brings together the ABS and BRES, and prior to 2009 it brought together the two parts of the Annual Business Inquiry (ABI).

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. Material inputs are deflated with the ONS producer price inflation data. The capital stock is deflated with the ONS gross fixed capital formation deflator.

A.1.2 Cleaning

For the purpose of our production function estimation, we exclude sectors: Agriculture, Public Sector, Finance & Insurance, Education, and Health.[9] 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]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
Capital Share	0.27	-	0.06	0.19	0.60	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 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.

C Model TFPR Decomposition

The analysis in this section supports using $TFPR$ as a proxy for A. We analyse A in our theory, but we measure $TFPR$, instead of A, in our empirical analysis. We show

that TFPR implies by our theory behaves similarly to A.

Revenue is the product of the inverse demand function $p_t(\jmath) = N_t^{\frac{1+\epsilon-\mu}{\mu}} \left(\frac{Y_t}{y_t(\jmath)}\right)^{\frac{\mu-1}{\mu}}$ and the production function $y_t(\jmath) = A_t(\jmath)^{1-\nu}\left[k_t(\jmath)^{\alpha}\ell_t(\jmath)^{1-\alpha}\right]^{\nu}$, thus:

$$p_t(\jmath)y_t(\jmath) = N_t^{\frac{1+\epsilon-\mu}{\mu}} Y_t^{\frac{\mu-1}{\mu}} y_t(\jmath)^{\frac{1}{\mu}} = N_t^{\frac{1+\epsilon-\mu}{\mu}} Y_t^{\frac{\mu-1}{\mu}} A_t(\jmath)^{\frac{1-\nu}{\mu}} \left[k_t(\jmath)^{\alpha}\ell_t(\jmath)^{1-\alpha}\right]^{\frac{\nu}{\mu}}. \tag{61}$$

Taking logs yields:

$$\ln p_t(\jmath)y_t(\jmath) = \alpha\frac{\nu}{\mu}\ln k_t(\jmath) + (1-\alpha)\frac{\nu}{\mu}\ln \ell_t(\jmath) + \frac{1-\nu}{\mu}\ln A_t(\jmath) + \frac{\mu-1}{\mu}\ln Y_t + \frac{1+\epsilon-\mu}{\mu}\ln N_t. \tag{62}$$

Therefore the revenue function residual is:

$$\ln TFPR_t(\jmath) = \frac{\nu}{\mu}(1-\alpha)\left[\ln \ell_t(\jmath) - \frac{\ell_t^{\text{tot}}(\jmath)}{\ell_t(\jmath)}\ln \ell_t^{\text{tot}}(\jmath)\right] + \frac{1-\nu}{\mu}\ln A_t(\jmath) + \frac{\mu-1}{\mu}\ln Y_t + \frac{1+\epsilon-\mu}{\mu}\ln N_t. \tag{63}$$

For small ϕ the first term is negligible. The $\frac{1-\nu}{\mu}\ln A_t(\jmath)$ 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 $p_t(\jmath)$ once it is purged of individual effects. In other words, it is the logarithm of $p_t(\jmath)y_t(\jmath)^{\frac{\mu-1}{\mu}} = N_t^{\frac{1+\epsilon-\mu}{\mu}} Y_t^{\frac{\mu-1}{\mu}}$. Our classification of technical efficiency and a demand shifter is equivalent to Decker, Haltiwanger, Jarmin, and Miranda (2020, p. 3, 961).

Special Cases: Under a standard CES aggregator, external returns to scale (love of variety) are $1 + \epsilon = \mu$. In this case N_t disappears from $\ln TFPR_t(\jmath)$. 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 $\frac{\mu-1}{\mu}\ln\frac{Y_t}{N_t}$.

C.1 TFPR Numerical Decomposition

Figure 11 plots the contribution of each component of $\bar{TFPR}$ in Equation (59). $\bar{TFPR}$ is declining in ν. The majority of the variation in $\bar{TFPR}$ is driven by the changes to $\bar{A}$. Figure 12 shows that the implications of changing ν are similar for true technology A and TFPR.

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

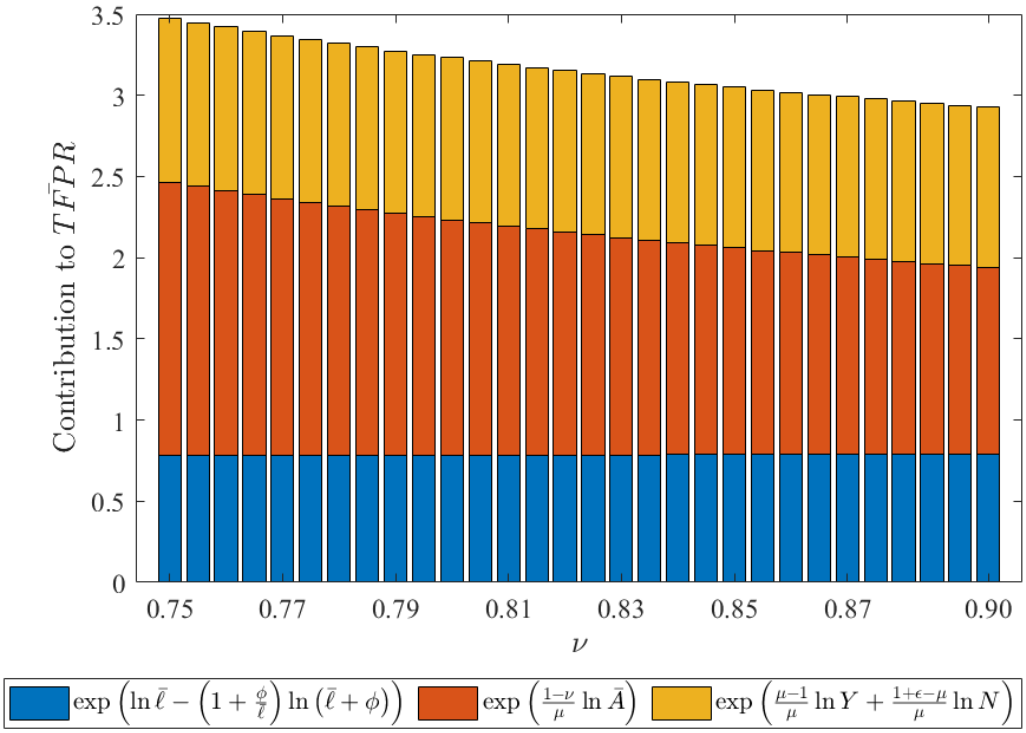
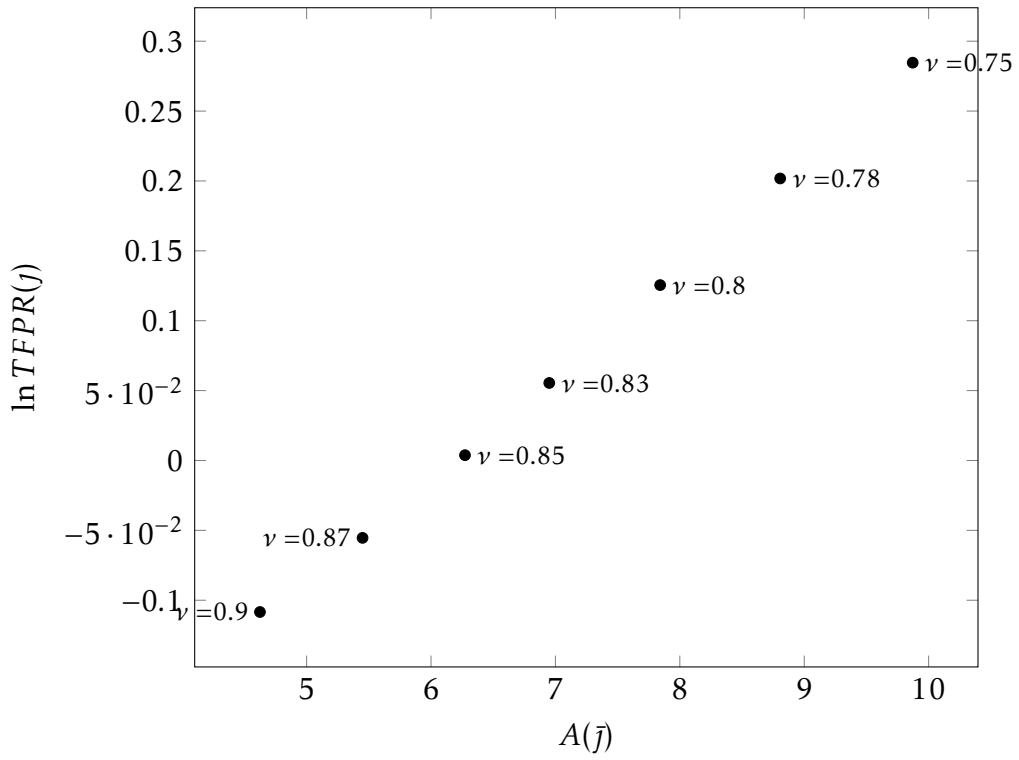


Figure 12: Industry equilibrium average technical efficiency and implied TFPR



D Further Figures & Tables

D.1 Additional Simulations

Figures 13 and 14 show results from the calibrated model simulation. We plot the number of operating firms N and entrants E as ν varies. The following follows from $N = \frac{1-u}{\phi}$ and $N = (1-J)E$.

Figure 13: ν and E

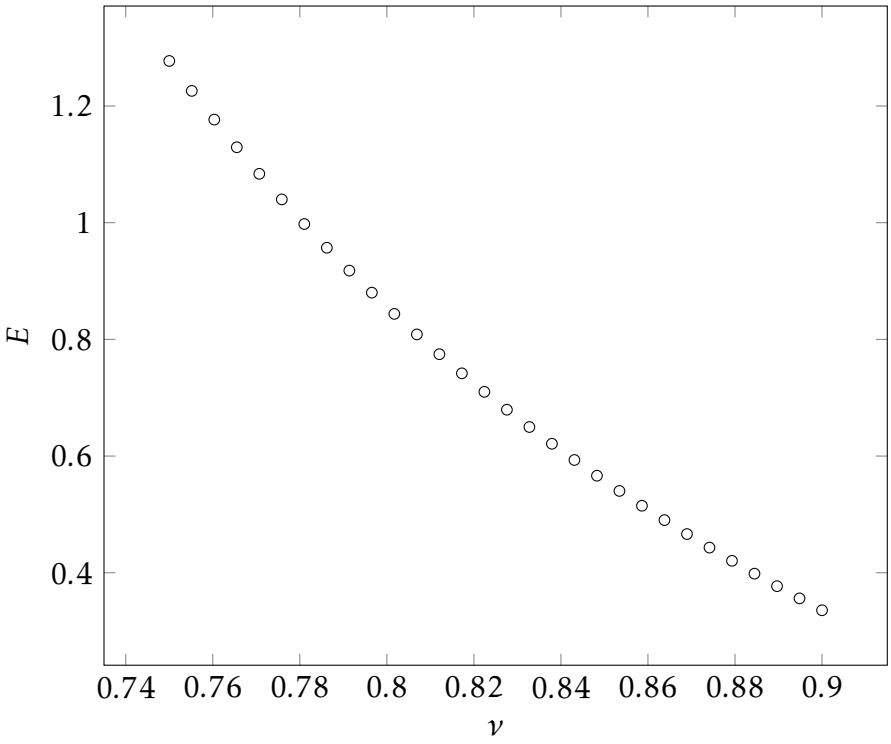
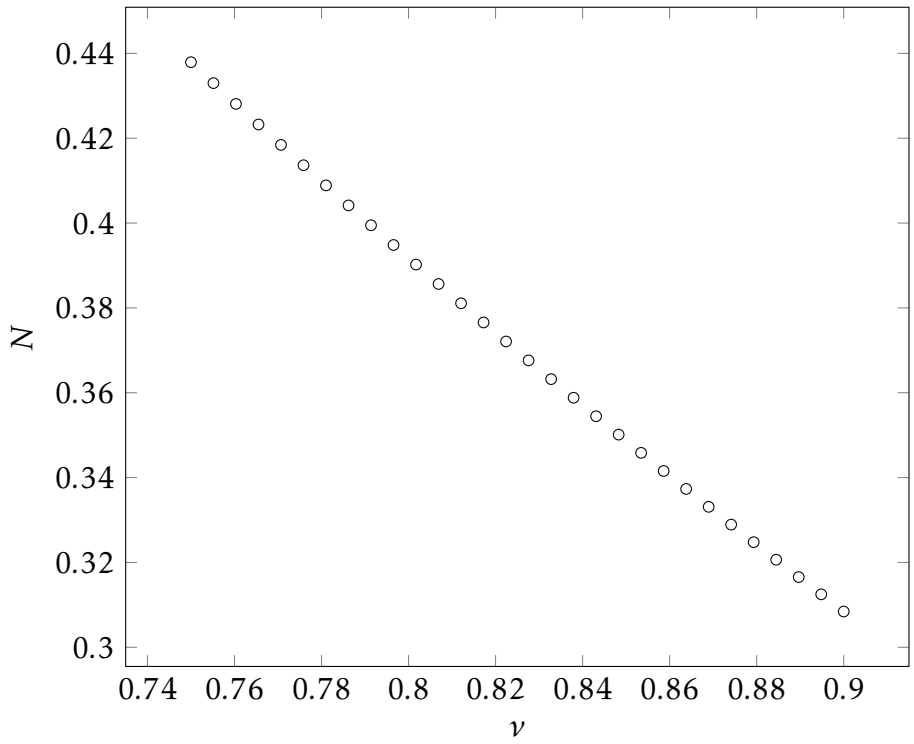


Figure 14: ν and N



D.2 Additional Testable Implications

D.2.1 Returns to Scale of Large and Small Firms

Table 13 shows the returns to scale estimates and summary statistics when we estimate production functions on sub-samples of large (top quintile of employment) and small (bottom quintile of employment) firms. The results confirm an important prediction of our model that large firms have lower returns to scale and smaller firms have higher returns to scale. This implicitly provides estimates of ν and μ since large firms returns to scale converge on ν and small firms returns to scale converge on μ.

Table 13: Returns to Scale for Top vs Bottom Employment Quintile

	RTS	$\ln TFPR$	L	N
Bottom Quintile	1.310	1.531	2.312	100,702
Top Quintile	0.967	3.573	1078.8	106,724

Firms are split into quintiles by employment. These are approximate due to integer constraints. The bottom quintile contains 18.8% of firms in the data, while the top contains 19.7% of firms.

D.3 Calibration Robustness

Table 14 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 14: 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.

D.4 Additional Results for GNR

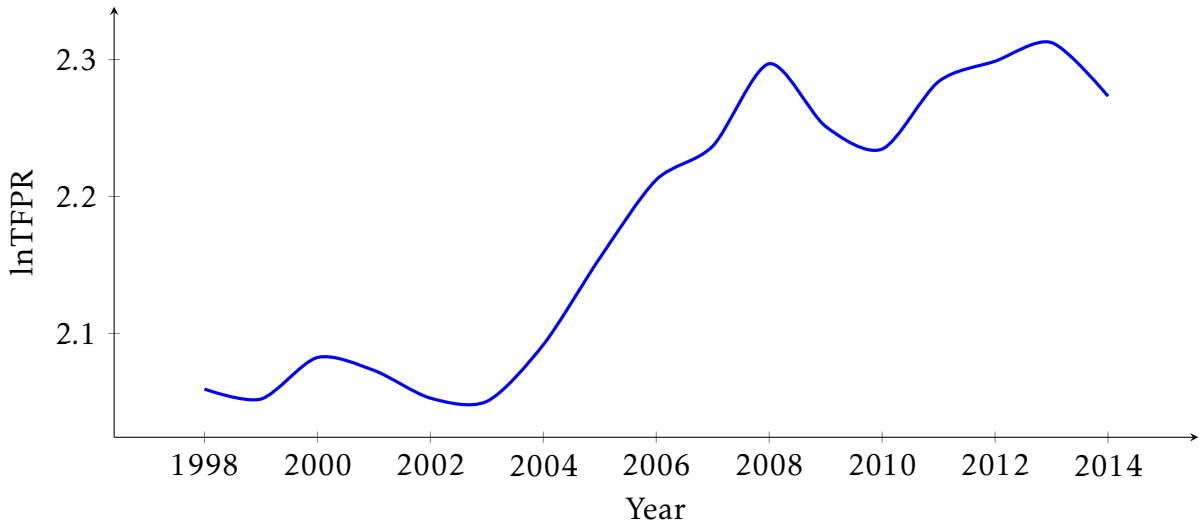
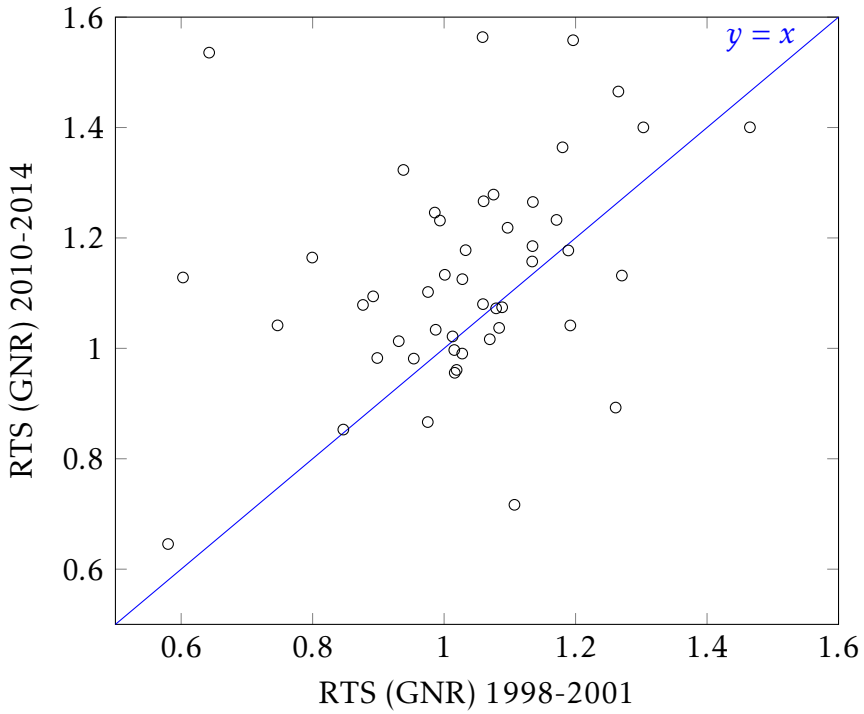


Figure 15: Aggregate TFPR (GNR)

Figure 16: Changing Returns to Scale by 2-digit SIC, GNR Estimation.



Comparison of returns to scale at 2-digit SIC level, from 1998 - 2001 to 2010 - 2014. Line is 45 degree line: points above that line are consistent with a rise in returns to scale.

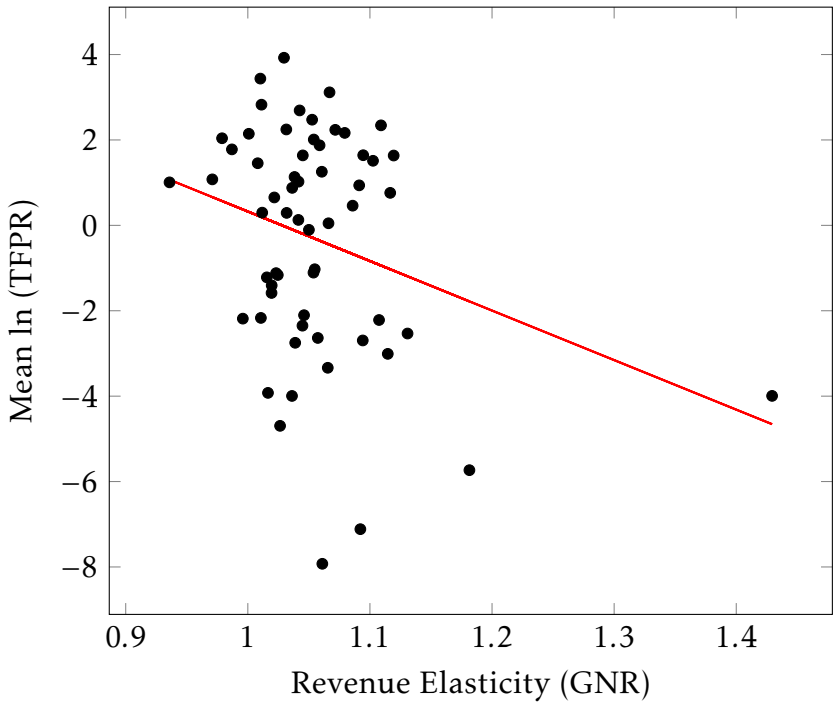


Figure 17: Relationship between Revenue Elasticity and TFP, GNR

E Pareto Distributed Productivity

We obtain a measure of productivity $A(\jmath)$ from a random draw on the unit interval $\jmath \in [0,1]$ using inverse transform sampling. 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.

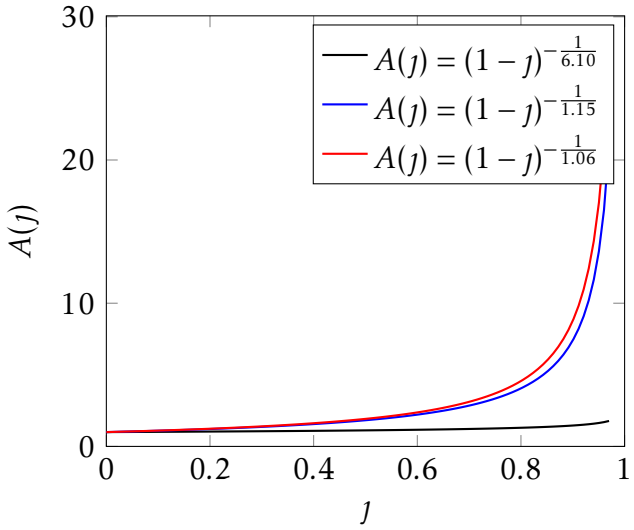


Figure 18: Productivity with Pareto Distribution, $h = 1, \vartheta = \{1.06, 1.15\}$