UNIVERSITY COLLEGE LONDON

An Examination of Imperfect Competition in Macroeconomics

Dajana Xhani

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in the

Department of Economics

August 2022

Declaration of Authorship

I, Dajana Xhani, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Dajana Xhani
August 4, 2022

Abstract

This thesis investigates ways in which more general notions of market power can be incorporated in macroeconomics and can be leveraged together with large firm-level datasets to study misallocation and policy interventions in new ways.

The first chapter studies the rise of markups in the US, with the objective to distinguish between the pessimistic view that markups have risen due to higher barriers to entry versus a more optimistic view that higher markups are the result of higher productivity. I answer this question in an oligopolistic framework, with multi-product firms that follow exogenous processes of productivity innovation and expansion into new markets. Matching the model to Compustat data for Manufacturing firms, I find that the increase in markups is driven by an increased dispersion in productivities with the expansion rate into new markets having increased. In the model, these forces are what generates a higher dispersion in markups driven by the right-tail of the distribution, with almost no change at the median or below.

The second chapter studies misallocation in the generalized monopolistic competition model with heterogenous firms. In particular, it offers a new welfare statistic that is derived from mapping firm-level elasticities into an aggregate effect. This allows the researcher to answer welfare question without making any parametric assumptions on the demand schedules and therefore offers a way to relax the functional form restrictions made in previous work on misallocation. I show that the total welfare change can be decomposed into three channels: (i) the direct effect of the shock, (ii) a selection effect that arises as the least productive firm in equilibrium changes and, (iii) a reallocation effect as production shifts across firms. I also pursue some extension of the baseline version

of the framework and confirm that the welfare formula remains tractable and intuitive.

The third chapter examines whether nonlinear firm taxes can have sizable welfare effects in light of the high markup dispersion (and thus potentially misallocation) that has already been documented in the literature. To make the sufficient statistic derived in the previous chapter operational, one needs estimates of both markups and output responsiveness at the firm-level. I propose a novel way to non-parametrically recover the responsiveness parameter by exploiting the monopolistic setup and some commonly used assumptions on the production side. I apply this estimator, together with the standard one on markups to a large dataset of UK firms. I conclude that it is welfare-improving to subsidise small firms at the expense of large ones and moreover, even a simple revenue-neutral two-tier sales tax can deliver substantial gains.

Impact Statement

The motivation for this work has been provided to a large extent by recent findings in the economic literature that suggest that market power, as measured by the firm's ability to charge prices above marginal costs has trended upwards in recent decades. The goal has been to extend our modelling in a way that can capture these new developments while also being parsimonious enough to serve as a framework for exploring more economic mechanisms in future work.

In academic terms, this work contributes to the literature in the following ways. Chapter 2 relates to the misallocation literature and offers a new sufficient statistic formula for monopolistic models with heterogenous firms. It highlights two important conceptual points. Firstly, and as has already pointed out by Bulow & Pfleiderer (1983), markups are not a sufficient statistic for welfare and thus the current empirical work on markups, while extremely illuminating and useful is also insufficient in this direction. Secondly, the notion of *average markup* that is needed for a second best constrained-optimum is not a cost(sales)-weighted mean but a different measure altogether that resembles a consumer surplus and cannot be inferred without further assumptions. Chapter 3 proposes a novel way of estimating output responsiveness at the firm level from the joint distribution of firm sales and variable costs. I apply this method to a large UK dataset, and together with a standard measure of markups I am also able to infer price passthrough at the firm level. These are often needed to calibrate macro models and thus can be useful to researchers that study business cycles, especially if they want to allow for firm heterogeneity or differences between sectors.

In terms of policy implications, this thesis offers some stark recommendations. The results in Chapter 1 suggest that the increase in

markups for a minority of firms can be replicated in a stochastic multi-product firm model by an increase in the productivity dispersion. This suggests that entry barriers per se are not the culprit for higher markups and profits. Instead, the innovation process, the patent system and all other aspects of the regulatory and tax environment that affect the diffusion of new technologies should be scrutinized more thoroughly. In Chapter 3 I take a more prescriptive approach and ask whether in light of the vast disparities between firm markups, governments can use taxation to improve consumer welfare. The empirical work concludes that there are non-trivial gains from charging higher tax rates to large firms and cutting taxes for smaller ones in a revenue-neutral way. These results challenge the orthodoxy that misallocation comes from small firms simply because they are less productive than their larger competitors.

Acknowledgements

The PhD has been a challenging experience, full of excitement, learning and intellectual stimulation as well as moments of exasperation or monotony. In the end, it has proved to be a worthwhile endeavor and I am very grateful to everyone that has helped me along the way.

I would firstly like to thank my two supervisors, Vincent Sterk and Victor Rios-Rull. I am immensely indebted to Vincent for all the time and support he has given me from day one. I often anticipated our regular meetings with trepidation but I appreciate now how incredibly fruitful his comments and advice were. I know he takes his role as mentor very seriously and like his research, he does an excellent job at it. Victor became my supervisor only a year ago, but he left an important mark nonetheless. I learned a lot from our interactions and I would especially like to thank him for helping me communicate my research in a clearer and more impactful way.

The research community at UCL is fantastic and too vast to enumerate but I would particularly like to mention the Macro group. Many thanks to Alan Olivi, Franck Portier, Morten Ravn, Ralph Luetticke and Wei Cui for not only being amazing researchers but also some of the nicest people to interact with. I have learned a lot from them and I am deeply grateful.

My peers have made the journey infinitely better and I am very glad to have met so many bright, ambitious, and friendly, caring graduates. Some have passed on their insights and experience as senior students, others have shared with me the anxieties of an exam or the first-ever presentation and for some I have hopefully been a helpful guide as they navigate the initial years of the PhD. So many thanks to Alex, Amir, Anna B., Anna V, Anusha, Chanwoo, Enrico, Fernanda, Francesca, Gherardo,

Gonzalo, Guillermo, Jihyun, Linda, Morgane, Silvia, Thomas, Uta, Vlad, and everyone else that I have met at UCL these past years.

Lastly, I must acknowledge the great indebtedness I owe to my family. Ultimately, this was only possible thanks to my mum and this thesis is dedicated to her.

Contents

List of Figures

List of Tables

Chapter 1

Decomposing the Rise in Markups: Technology or Competition?

1.1 Introduction

A host of different empirical findings suggest that the US economy has become less competitive in the last decades. While the interpretation of these economic trends is still being debated, there are at least three facts that are now more or less well-established. Firstly, there has been a well documented increase in average markups in the US (De Loecker & Eeckhout (2017),Hall (2018)). Secondly, concentration has increased in most industries and is associated with a lower labour share (Autor et al. (2017)) and lower investment (Gutiérrez & Philippon (2017)). Thirdly, the TFP dispersion of firms operating in the same industries has increased (Decker et al. (2018)).

The last two facts suggest two tentative and differing explanations for the rise in markups. The larger market share captured by the biggest firms may indicate higher barriers to entry and/or a laxer regulatory environment. For example, a recent study by Blonigen & Pierce (2016) concludes that M&A activity is associated with an average increase in markups across U.S. manufacturing industries while there is little evidence of productivity improvements either at the plant or firm level. On the other hand, the greater dispersion in productivity suggests that the

increase in markups might be driven by 'superstar firms', highly productive firms that can charge high markups while also keeping prices low. This explanation has been favoured by Van Reenen who notes that "*the success of such firms may be as much due to intensified competition for the market rather than anti-competitive mergers or collusion in the market*".

While the second explanation presents a more positive view of the underlying changes, it still remains the case that the slowdown in technological diffusion or a higher cost of imitation by lagging firms have a negative impact on aggregate output. Furthermore, the policy implications are quite different under these two scenarios so it is important to discern between the two.

An important first step in this research agenda is to understand the 'mechanical' causes that have given rise to the observed change in markups. Standard macroeconomic models are not well-adopted for such a task since the usual CES assumption on the utility function implies that markups are the same for all firms.

In order to generate a distribution in firm-level markups, I follow an approach adapted from the growth literature. Specifically, I assume that there is a finite number of competitors producing undifferentiated varieties of a good and they interact strategically with each-other. To allow for a richer set of market structures I replace the commonly used assumption of limit pricing with one of Cournot competition. The main distinction between these two frameworks is that under limit pricing, productivity improvements by the incumbent firm in the market have exactly zero pass-through to prices. With Cournot on the other hand, productivity changes by any of the firms will matter for determining the good's price and one can show that from the consumer's perspective, productivity improvements by laggard firms lead to larger welfare gains than improvements of highly productive firms.

On top of that, I allow for firms to be multi-product entities, where expansion into new markets follows an exogenous Poisson process. The productivity process is product specific and happens independently across and within firms. This implies that over time, a firm that is active in a given market can become so productive so as to price out unproductive

laggards in that market. Furthermore, the assumption that a firm starts at the lowest level of productivity when entering a new market gives rise to endogenous entry barriers. In other words, some markets that are dominated by one or a few highly productive firms are effectively closed down to new competitors because the price is too low for any new firms to want to join.[1]

To solve for the equilibrium steady state, one needs to keep track both of the vector of productivities of competing firms in each market, as well as each firms vector of good-level productivities, prices and market share so as to compute the firm-level statistics that can be taken to the data. Numerically, this is a challenging problem which I simplify by making use of sparse matrices together with an assumption that the firm's productivity of producing a given good can only take a finite number of values.

I calibrate the steady state of the model using Compustat data on Manufacturing firms for the years 1980 and 2010. I find that changes in the productivity process account for the increase in firm-level markups which is driven by the top of the distribution. I find that productivity improvements upon innovation increase by around 15 percentage points while the probability of innovating in any given product goes down from about 10% to 7.6%. This leads to an increased dispersion in the productivity distribution of firms that can account for the skewed increase in markups. The exogenous entry into new markets also goes up which suggests that the barriers to entry story is not what is driving the dispersion in firm markups.

These results also have some interesting implications for the competitive structure of goods markets. In particular, the model suggests that the distribution of the number of competitors in each market has much fatter tails in 2000 compared to 1980. The share of markets with only 2 or 3 competitors has increased by almost 20pp but there has also been

[1]Both of these forms of endogenous exit and endogenous barriers to entry are purely the result of strategic competition in a static game. Although dynamic oligopoly would be a natural extension, it can usually only be solved using numerical methods which would render the solution of a general equilibrium model much less feasible.

an increase in the share of markets with 16 or more competitors. Such technological changes can also explain a more dispersed distribution of firm sales. In welfare terms, these changes are overall positive and they imply an output increase of 13.5%. The fall in the innovation rate is the only change that weighs down on welfare and I estimate that had it remained unchanged at its previous level, output would have increased by a further 10pp by 2000. This is very substantial and suggests that studying the endogenous mechanisms that determine firm productivity is of great economic importance.

Related Literature

This paper relates to the empirical work on firm markups and profits. De Loecker & Eeckhout (2017) show that average markups of Compustat firms have increased from about 18% in the 1980s to 67% in the late 2010s. Hall (2018) uses industry-level data and an instrumental variable approach and also finds an upward trend, albeit of a smaller magnitude. Barkai (2016) provides a decomposition of aggregate accounts data between labour, capital and profit share to come to a similar conclusion.

Importantly, De Loecker & Eeckhout (2017) have shown that most of the action in markups comes from the upper tail of the distribution, while the median and lower quantiles have barely moved. This is a crucial finding that is corroborated by other studies. Kehrig & Vincent (2017) use Census data on manufacturing firms and find that the overall fall in the labour share[2] has increased not due to a general fall across establishments, but because establishments with low labour shares have increased in size.

The model builds upon the framework in the endogenous growth literature started by Aghion & Howitt (1990) and Grossman & Helpman (1991), with more recent examples including Klette & Kortum (2004), Peters (2012) and Perla (2015). The crucial distinction is that I do not

[2]Assuming that labour can be freely adjusted, we have that the inverse of the labour share is proportional to markups under the cost-minimization condition of the firm (De Loecker & Warzynski (2012)). Furthermore, if we assume that the elasticity of output to labour has not changed, we can interpret changes in the labour share as changes in markup.

endogenize innovation at the firm level in the form of an investment choice but assume that the rate is exogenously given. Nonetheless, the amount of innovation that takes place depends on the equilibrium distribution of market structures since firms can only innovate in product lines that they are currently producing. This competition effect shows up because I replace the limit pricing assumption used in the growth literature with a rich model of quantity competition à la Cournot. This framework has been used by Atkeson & Burstein (2008) to study markups and price pass-through in international trade, but they abstract from firm-level productivity growth and multi-product firms.

This paper proceeds as follows. In Section 1.2 I present the model. Section 1.3 discusses the data, the calibration strategy and presents the result. Finally, Section 1.4 concludes.

1.2 Model

Time is discrete. There is a continuum of identical individuals that supply their unit time inelastically at the prevailing wage W. They own the firms and receive the totality of profits (Π) that firms make. Their preferences are given by a CES aggregator over individual goods m and they solve a static maximization problem given by

$$\max_{[q_m]\geq 0} \left(\int_{m\in\mathcal{M}} q_m^{\frac{\eta}{\eta-1}} \, dm \right)^{\frac{\eta-1}{\eta}} \quad \textit{subject to} \quad \int p_m q_m \, dm \leq W + \Pi. \quad (1.1)$$

Every good m is supplied by a finite number of firms, which compete a la Cournot to determine a unique price for that good (market). I will hereafter use the words good and market interchangeably. [3]

Discussion. The mass of goods is an exogenous variable in this model but it has important implications for competition and welfare. One can think of at least two recent trends that would have opposing effects on the number of goods on offer. On one hand, the increased availability of consumer data implies that firms are better able to make products catered

[3] This is to allow both for the interpretation of goods as distinct products, but also as the same product being offered at different geographies (markets) by a potentially different set of firms.

to finer categories of consumers. Furthermore, technological advances have made it cheaper to customize certain products to the individual level and this is one way in which firms can extract more value from consumers, especially if there are only a few big firms that can offer this. On the other hand, the rise of internet usage and online shopping means that consumers can learn about products of which they were previously unaware of at a much faster rate. This implies that they potentially have access to more substitutes than before thus reducing the segregation of markets for the same goods.

1.2.1 Market Structure

The competition structure follows closely the double CES model firstly used by Atkeson & Burstein (2008) with the important restriction that goods within a market are perfect substitutes in my setup. This assumption is necessary for getting closed-form solutions of price and market shares as function of productivities. Firms have a linear production technology with labour being their only input. Let z_{im} denote the productivity of firm i in producing good m, q_{im} be its quantity and P_m be the price of the good. The firm's profit maximization problem is

$$\max_{q_{im}} \left(P_m - \frac{W}{z_{im}} \right) q_{im}, \tag{1.2}$$

subject to the inverse demand function for good m

$$P_m = D \times \left(\sum_{j \in N_m} q_{jm} \right)^{-1/\eta}. \tag{1.3}$$

The *demand index* D is a function of aggregate consumption C and the aggregate price level P only, with the formula given by $D = C^{1/\eta} P$. Competition is strategic because firms internalize the impact that their choice of output will have on the price of the good but they take the demand index D as given as markets are atomistic.

I solve for the equilibrium vector of output by using the FOCs of the firms' maximization problem given by

$$P_m \left[1 - \frac{1}{\eta} \frac{q_{im}}{q_m} \right] = mc_{im}. \tag{1.4}$$

This equation also clarifies why perfect substitutability of goods is critical for getting closed-form solutions by using the aggregation constraint that $q_m = \sum_{j \in N_m} q_{jm}$, ie. total market output is simply a sum of individual firms' output. Hence, we first solve for the unique good's price

$$P_m = \left(1 - \frac{1}{\eta n}\right)^{-1} \overline{mc}. \tag{1.5}$$

In words, equation 1.5 says that the price of a good is a markup on the average unweighted marginal cost of all producing firms. The level of industry markup is in turn determined by the degree of substitutability between goods (η) and the number of competitors. This second channel is extremely important because it clarifies how low entry will results in lower welfare. Substitute this back in equation 1.4 to recover the formula for market each firm's market share

$$s_{im} = \eta \left[1 - \left(1 - \frac{1}{\eta n}\right) \frac{mc_{im}}{\overline{mc}}\right]. \tag{1.6}$$

Unsurprisingly, more productive firms will have higher market shares and higher markups, with the later given by

$$\mu_{im} = \left(1 - \frac{s_{im}}{\eta}\right)^{-1}. \tag{1.7}$$

Iterative Solution. Although the model has a closed-form solution, perfect substitutability implies that there are cases in which some firms will choose to produce zero output and so they need to be excluded from the function that determines prices. Linearity of costs implies that for any vector of potential producers there is going to be a cutoff marginal cost above which firms will choose not to produce. To find this cutoff, I calculate the price and the market share for the full vector of firms. I then discard firms for which the market share is negative and continue this process iteratively until I have a set of firms with strictly positive market shares. This method is guaranteed to find the Nash Equilibrium of the game and most importantly that equilibrium is unique even when firms are allowed to tie in the marginal cost ranking. In other words, if there are two or more firms with the same marginal cost, there is no Nash Equilibrium that sustains only a subset of these firms producing so

that either all tying firms will produce zero or they will all produce some strictly positive amount.

Discussion. Notice that although the model has perfect substitutability between producers of the same good, it still captures some important channels of competition. In particular, the pass-through rate is always positive and will be higher for smaller producers. This is in line with the empirical findings of Amiti et al. (2019) but cannot be reproduced in all strategic models of competition. For example, with prefect substitutability and Bertrand competition, choke pricing implies that the market leader will never pass on any cost reductions to consumers.

1.2.2 Description of Firms

Firms are all identical ex-ante, but they face a stochastic growth process which implies that the distribution of firms will be non-degenerate. A firm in this economy is defined by a set of vectors with the productivities of all competitors in any given market where the firm is 'active'. A firm is said to be active in a given market if it produces a strictly positive amount. Then we have that the set $\{\{z_{m_1,1}, \cdots z_{m_1,n_1}\} \quad \forall m \in \mathcal{M}_i\}$ fully describes the state variables of the firm.

There are two ways for a firm to grow in this economy. Firstly, firms could experience a productivity increase for any of the goods that they are already producing, which other thing being equal will lead to a higher market share and markup for that good. Secondly, firms can grow their portfolio of goods by expanding into new markets. I will now describe these two processes in more detail. Following Grossman & Helpman (1991), the productivity process is modeled on a quality ladder with parameter θ, and the set of potential productivities is given by $\{1, \theta, \theta^2, \theta^3, \dots\}$. The probability of moving up the ladder is given by γ and is independent of the current productivity level and is iid both within and across firms. Note that firms never experience negative productivity

[3]Shutting down adjustments on the extensive margin, a first order approximation for market price is $\Delta ln(P_m) \approx \frac{mc_j}{\sum_i mc_i} \Delta ln(mc_j)$ implying that for the same percentage change in marginal costs, price will fall more if that happens to a smaller, i.e. higher marginal cost firm.

shocks in this model and the productivity jump can only be to the level above.

To eliminate the possibility that a firm or group of firms become so efficient in producing a good that they can never be challenged by new entrants, I assume that there are exogenous death shocks at the product level. The probability of a death shock is δ and these shocks are uncorrelated across competitors or within firms.

Expansion on the other hand will follow an exogenous Poisson process. Every period, firms draw from a Poisson process with probability λ, which determines the number of new markets in which they can enter.[4] Additionally, I impose the restriction that when a firm enters into a new market, it does so at the lowest level of productivity given by 1. This implies that firms cannot necessarily enter in all of the new markets that they randomly draw, given that some of these markets have a price that is too low for potential entrants.

Figure 1.1 illustrates the dynamic evolution of a single market in this economy by plotting the number of firms that are active in that market and the total amount of quantity produced, which is determined endogenously by the distribution of productivities.

[4]Every market has an equal probability of being entered by the firm which means that the draw could potentially be one of the markets in which the firm is already active. Since the mass of the firm is zero relative to that of all markets, this possibility will be ignored in practice.

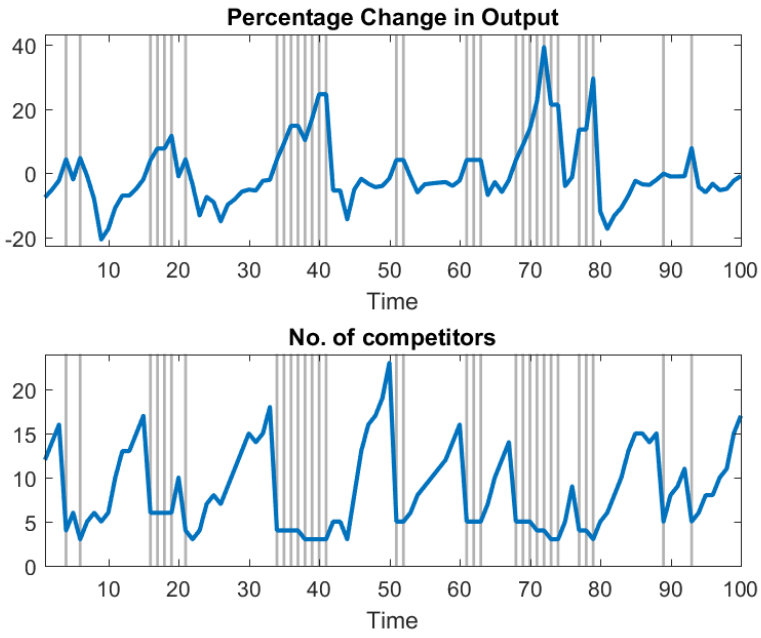


Figure 1.1: Simulated Market Outcome for a randomly chosen market.

1.2.3 Firm-level markups

Because our data is at the firm level only, we need to derive firm markups from the market-level markups that we have solved for above. One approach is to follow the cost-based estimator derived in De Loecker & Warzynski (2012) which is given by the elasticity of output times the inverse of the sales share for any variable input ($\theta_{input} \cdot \frac{Sales}{Costs_{input}}$). In our setup, the only input in production is labour and the elasticity is exactly equal to one.

In the main calibration, I will assume that the elasticity of substitutions across goods is one which implies that the expenditure of the household on each good is fixed and denoted by some constant K. The expression for total variable costs (TVC) can be derived by adding up

across all the goods that the firm produces

$$\begin{aligned} TVC &= \sum_{m \in \mathcal{M}_i} W \frac{q_{im}}{z_{im}} = \sum_{m \in \mathcal{M}_i} W \frac{s_{im} Q_m}{z_{im}} \\ &= \sum_{m \in \mathcal{M}_i} s_{im} \left(\frac{W Q_m}{K z_{im}} \right) K \\ &= \sum_{m \in \mathcal{M}_i} s_{im} (1 - s_{im}) K \\ &= K \sum_{m \in \mathcal{M}_i} s_{im} \left(\frac{1}{\mu_{im}} \right). \end{aligned} \tag{1.8}$$

Similarly, one can write down the expression for total firm sales as

$$\begin{aligned} \text{Sales} &= \sum_{m \in \mathcal{M}_i} p_m q_{im} \\ &= \sum_{m \in \mathcal{M}_i} \left(\frac{K}{Q_m} \right) q_{im} \\ &= K \sum_{m \in \mathcal{M}_i} s_{im}. \end{aligned} \tag{1.9}$$

Putting these together allows me to recover the model's counterpart for the sales to variable costs ratio

$$\mu_i = 1 \cdot \frac{Sales_i}{TVC_i} = \frac{\sum_{m \in \mathcal{M}_i} s_{im}}{\sum_{m \in \mathcal{M}_i} s_{im} \left(\frac{1}{\mu_{im}} \right)}. \tag{1.10}$$

In words, the firm-level markup is simply the shares weighted harmonic mean of product-level markups, where the shares are constructed from the product sales of the goods.

Finally, note that one issue that arises when we set the elasticity of substitution across goods to 1 is that the problem of the firm is not well defined when in faces no competitors. One can see from equation (1.5) that the firm would want to set an infinitely high price and produce nothing. One way to deal with this case is to impose a regulatory constraint that puts a minimum level $q_m in$ on the quantity supplied. Quantitatively, how high one sets this amount is inconsequential for the models results as very few markets will be pure monopolies in equilibrium.

1.3 Calibration

I solve for the model by simulated method of moments to find the steady state distribution of firm level variables. Although there has been some

interesting work that documents a substantial fall in start-up rates in the US since 1980 (see for example Pugsley & Şahin (2015)), I sidetrack this issue and normalize the mass of firms to equal 1.

This leaves me with the ratio of markets to firms as the free parameter. I also use the identity $\int_{i\in\mathcal{F}} \text{Sales}_i di = K \times m_{\mathcal{M}}$ to solve for the market expenditure directly for any given ratio of markets to firms and the average firm sales that I get from the data. I use the MCMC (Laplace Type Estimator) developed by Chernozhukov & Hong (2003) to estimate the 5 remaining parameters consisting of the ratio of markets to firms, the Poisson entry rate, the probability of exit, the probability of innovation and the productivity gap ($\{m, \lambda, \delta, \gamma, \theta\}$).

1.3.1 Data

The calibration is based on Compustat data, which includes all public companies in the US. Because there is not an obvious definition of industry in my model and firms expand randomly in the product/market space I choose to focus only on Manufacturing. Another reason for this choice is that the cost of producing manufactured goods is better defined than the cost of say service goods. This makes the analysis less general of course but it also makes it more robust to criticism that the COGS (Cost of Goods Sold) variable in Compustat in not a good measure of production cost as highlighted by Traina (2018)). As mentioned before, I solve for the stationary distribution only and hence need to pick a starting and end point. I choose the years 1980 and 2000 for my comparative statics exercise. The choice of the start year is relatively obvious given that a lot of papers emphasize the 1980s as a turning point, while for the end period I have chosen the year 2000 for two main reasons. On one hand, all of the important trends are already discernible by then and on the other hand by not pushing it further out I avoid some potential confounding factors like China joining the WTO or the build-up and impact of the Great Financial Crisis.

1.3.2 Results

Tables 1.1 and 1.2 show that the model fares relatively well in replicating the empirical moments for both periods. There are two important points to be made however. Firstly, the model cannot quite match the lower part of the markup distribution and for both periods the fitted lower markup quartiles are too high. This might not be too surprising given that firms are ex-ante endogenous and their productivity levels across markets are uncorrelated so that there is in a sense an averaging out of productivity differences (as well as competition intensity variations) at the firm level. This is a force that restricts how much different firms' outcomes are even in the presence of stochastic shocks. However, the model can still generate the widening gap in markups which is a crucial feature of the data. Secondly, the model does not capture the full extent of the increase in the profit share that we observe in the data. With a bit of rearrangement, one can rewrite the profit share in the model as

$$\Pi = 1 - \int_{i \in \mathcal{F}} \mu_i \tilde{s}_i \, di, \tag{1.11}$$

where $\tilde{s}_i$ stands for the firm level share in total sales. So in a sense, I cannot get quite enough sales mass to shift to firms with high markups.[5]

Table 1.1: Model Fit for 1980

	Moments	Data	Model
1	IQR of Markup change	0.053	0.055
2	Profit Share	0.384	0.349
3	Lower Quartile of Markup	1.234	1.429
4	Upper Quartile of Markup	1.552	1.657
5	Mean Markup	1.512	1.525

[5]One potential extension of the model is to allow for some non-zero correlation of productivity shocks at the firm level. This could help generate a greater level of production re-allocation and it could also be thought of as a measure of 'scalability' which is one other explanation that has been put forward in explaining the rise in markups.

Table 1.2: Model Fit for 2000

	Moments	Data	Model
1	IQR of Markup change	0.098	0.094
2	Profit Share	0.557	0.417
3	Lower Quartile of Markup	1.265	1.473
4	Upper Quartile of Markup	1.954	2.048
5	Mean Markup	1.431	1.431

Table 1.3 contains the estimated parameter values for the two periods of interest. I find that there has been a decrease in the ratio of markets to firms by a factor of three. This means that the prize of 'capturing' any particular market has increased. I also find that the probability of innovation has gone down from about 10% to 7.6%. On the other hand, the productivity gap conditional on successfully innovating has gone up by about 12.2%. The exogenous probability of exiting has also gone down by about 2 percentage points. Finally, although Table 1.3 shows that the absolute entry rate (λ) has gone down by about a third, the fall in the ratio of markets to firms has been even larger, implying that entry rate per market is actually higher in 2000 compared to 1980.

Table 1.3: Estimated Parameters

Parameter	Description	1980	2000
m	Ratio of markets to firms	3.318	1.024
λ	Poisson Entry Mean	9.244	6.069
δ	Exogenous Prob. of Exit	0.100	0.083
γ	Exogenous Prob. of Innovation	0.101	0.076
θ	Productivity Gap	1.258	1.414

One important underlying change that the model predicts is a polarization of product markets. Figure 1.2 plots the distribution of markets by number of competitors for 1980 and 2000. We see that in the second period there has been a hollowing out of the middle part of the distribution with much more mass on markets with only 2 or 3 competitors (which cumulatively account for more than half of all markets in the second period) but also more mass in the right tail of the distribution,

which includes markets with 15 or more competitors. The latter change is a consequence of the fall in both the exogenous probability of exit and innovation, as well as the increase in entry rate.

To understand the increase in markets with only a couple of firms, it is important to highlight the role that endogenous exit plays. I find that while in the first period the number of endogenous exits were about 3.5 times as much as the exogenous ones, that ratio has jumped to almost 10 for the latter period. Intuitively, this can be explained by the higher productivity gap that I calibrate for 2000. Consequently, when one or a few of the leading firms are hit by a positive productivity shock, it is much less likely for the laggards to survive in 2000 compared to 1980 thus generating higher endogenous exit rates.

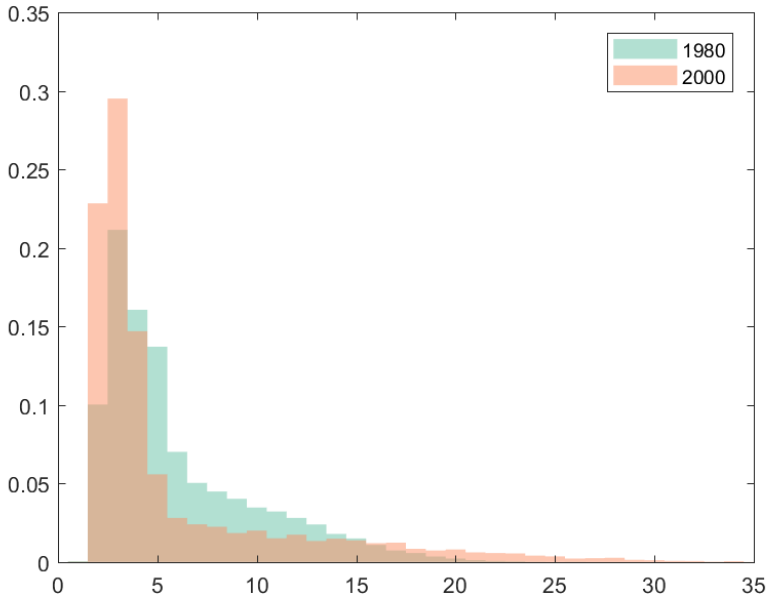
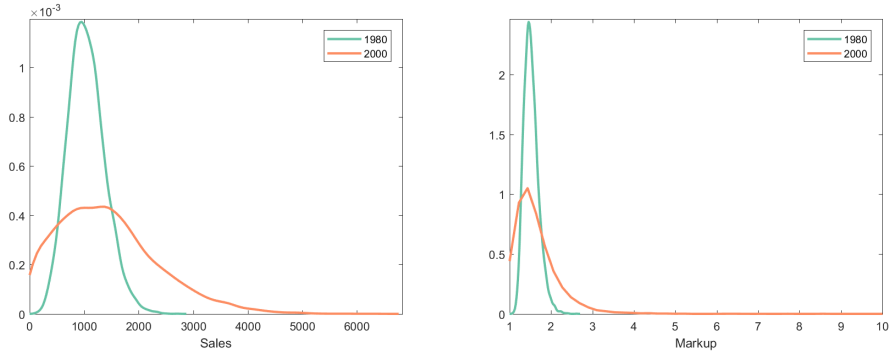


Figure 1.2: The distribution of the number of competing firms in each market.

Figure 1.3 plots the underlying distribution of firm sales and markups in the steady state for 1980 and 2000. While it is expected to see that the variance of the markup distribution has increased given that we were targeting the moments of this distribution, it is interesting to note that the same has happened for the size distribution of firms. In particular, not only do we get a much more pronounced right tail of the size distribu-

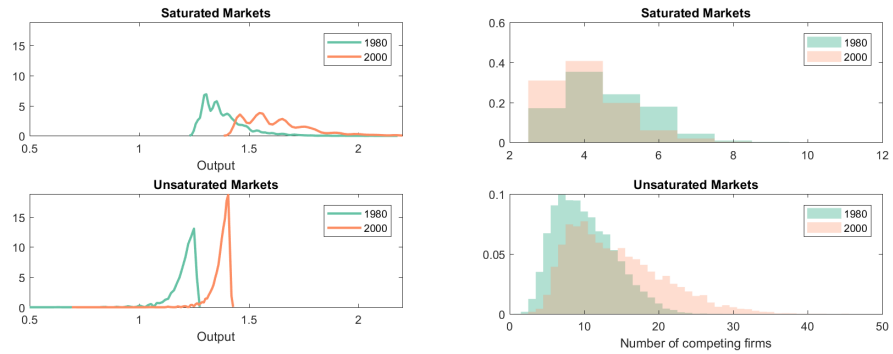
tion but also more mass on the smallest of firms. Note that to construct the distribution of firms for the latter period, I am keeping expenditure per market fixed at the 1980 value. Another option would be to deflate firm level sales by some price index so as to be able to express the last period sales in terms of 1980 dollars.



(a) Kernel Distribution of Firm Sales (b) Kernel Distribution of Firm Markups

Figure 1.3: Comparing the Steady State Distribution of Firms in 1980 and 2000

One informative way to partition markets is by looking at whether new firms can enter or not. I will call a market saturated if the current productivity levels of incumbents are such that an entering firm at the lowest level of productivity would optimally choose to produce nothing and hence exit. Notice that saturated markets are in general more productive and thus have lower prices (and hence higher quantities) than unsaturated markets.



(a) Histogram of the Number of Competitors (b) Kernel Distribution of Market Output

Figure 1.4: Comparing Saturated and Unsaturated Markets in 2000

1.3.3 Welfare Implications

I calculate that the estimated change in parameters is associated with a 13.5% increase in total output. Furthermore, using the steady state equilibrium, we can characterized the full distribution of market output which I have plotted in Figure 1.5.

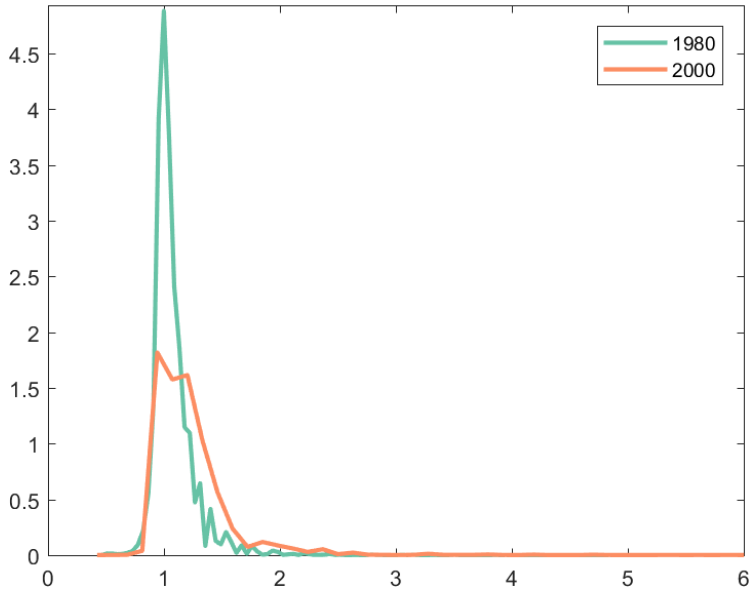


Figure 1.5: Kernel distribution of Market Output.

Another interesting point is to understand how the shift in each parameter has contributed to the total welfare impact. To this end, I conduct a counterfactual exercise where I take the parameter values of 1980 and solve for a new equilibrium by separately changing each parameter value at a time. The only exception is the Market Ratio and Poisson Entry parameter which I change jointly since although there is an absolute fall in the entry parameter, the fall in the ratio of markets to firms implies that the average entry per market has actually gone up. I find that all changes have had a positive impact on output with the exception of the falling probability of innovation. More specifically, I find that the increase in the productivity gap has had the largest impact of all and accounts for almost 10 percentage points of the total output increase.

Table 1.4: Decomposing Output Changes

Parameters	Output Change
Market Ratio and Entry	3.74%
Probability of Exit	3.62%
Probability of Innovation	-3.35%
Productivity Gap	9.91%
Total	13.9%

It is unsurprising then that the productivity gap increase is also the main driver of the observed change in moments. The following table presents the moments that we would observe in the model if we fix the parameter values at their old 1980 values and only change the productivity gap parameter to its new value.

Table 1.5: Comparing the Fit of the model in 2000

	Moments	Data	Model	Counterfactual
1	IQR of Markup change	0.098	0.094	0.083**
2	Profit Share	0.557	0.417	0.427
3	Lower Quartile of Markup	1.265	1.473	1.586**
4	Upper Quartile of Markup	1.954	2.048	1.947
5	Mean Markup	1.431	1.431	1.431

We see that although the productivity gap parameter generally does a very good job by its own, it fails to account for the magnitude of the increase in the dispersion of yearly markup changes as well as implying that the lower quartile markup is even higher. This latter fact is somewhat predictable given that without a fall in the probability of innovation, the model cannot generate the kind of 'winner-takes-all' dynamics that explain the increased dispersion in markups in the full model.

1.4 Conclusion

This chapter studies a model of multi-product firms with exogenous innovation, entry and exit processes and Cournot competition at the market level. The strategic interaction between firms gives rise to an endogenous

channel of entry and exit as firms that are too unproductive relative to their competitors will fail to enter or survive in a market. These forces are important for aggregate productivity as the firms efficiency at making a good can only improve over time as it produces that good. In that sense, this set-up captures some aspects an endogenous learning-by-doing dynamic that depends on the degree of competition.

I take the model to the data by calibrating to 1980 and 2000 using the Compustat data for Manufacturing firms. I conclude that there has been a fall in the innovation probability and a large increase on the productivity improvement if innovation happens. Coupled with the fact that highly productive firms can push their competitors out of the market by driving the price down, this points to an intensification of a 'winner-takes-all' economy. I find that these changes result in a fall in the number of competing firms in the average market, that is mostly driven by the increase in productivity differences.

This implies that although some firms are making very high markups, their underlying productivities are large enough to sustain low prices and high output. This does not mean that the picture is all rosy however. I calculate that if the rate of innovation had remained unchanged from its previous level, output would have increased by 23.5% instead of 13.5%. In policy terms, my findings support interventions that accelerate innovation diffusion and faster technological adoption rather than subsidizing entry of new firms.

Chapter 2

A Sufficient Statistic in General Equilibrium Monopolistic Models

2.1 Introduction

Monopolistic competition with constant demand elasticity is one of the building blocks of modern microfounded models that are employed in fields as diverse as growth, trade and the study of nominal rigidities and business cycles. This framework has proved to be parsimonious enough to study many different economic mechanisms, however in its standard Dixit & Stiglitz (1977) CES formulation it fails to generate a distribution of markups. This shortcoming can be corrected by assuming some other parametric utility form and there is a sizable and growing number of papers that do so, for example using linear demand Melitz & Ottaviano (2008), translog Bilbiie et al. (2012), QMOR Feenstra (2018) or CREMR Mrázová et al. (2021). This list, which is by no means exhaustive, displays the wide range of functional forms that have been developed and used in previous work and demonstrates that there is no consensus on how to deviate from the constant elasticity benchmark. [1]

[1] An expanding set of demand schedules that have attractive theoretical properties or can capture certain features of the data is in itself a highly valuable resource. The concern is whether sometimes, the specific parametric assumption rather than the empirical evidence drives the results when it comes to questions like misallocation.

To deal with the multitude of demand curves, this chapter proposes a sufficient statistic formula that is valid for all monopolistic models with heterogenous firms, and can be used to answer some interesting economic questions. For example, one might want to know what would be the welfare effects of a technology shock that hits only the 5% most productive firms in the economy or in a particular sector. On the other hand, given the current debate on the rise of markups, one might contemplate using differential taxation of firms akin to that on personal income as a means to alleviate misallocation and increase welfare. The sufficient statistic approach adopted in this paper can answer these question without imposing further and potentially arbitrary assumptions on the functional form of utility or the particular distribution of firm productivities.

The welfare formula presented in this chapter maps firm-level micro elasticities into a macro outcome taking into account general equilibrium effects. I show that the total welfare change can be decomposed into three channels: (i) the direct effect of the shock, (ii) a selection effect that arises as the least productive firm in equilibrium changes and, (iii) a reallocation effect as production shifts across firms. The reallocation channel features two key firm-level statistics: the markup (μ_f) and the output responsiveness (Δ). Markup is the usual price to marginal cost ratio and therefore is informative of the utility gains from consuming an extra unit of that particular variety. Output responsiveness is defined as the equilibrium percentage change in output for a given percentage shock to production costs. Intuitively, this matters for welfare because output changes multiply per unit utility gains as summarized by the markup. Moreover, the aggregate statistic needed for weighing selection and reallocation in this formula is a consumer surplus like measure ($\mathcal{M}$) and not a markup average.

The framework is based on the generalized monopolistic competition model with heterogenous firms that produce different varieties of the same good. The utility function is symmetric and additive across varieties but otherwise left unrestricted. All firms produce using the same technology function up to a Hicks-Neutral productivity term which maps into level differences in production costs. Profit maximization implies that more productive firms will be larger as lower variable costs drive

them down the demand schedule. How markups vary with firm size is determined by how the elasticity of demand changes along the demand schedule and is completely unrestricted in this setup. [2] To account for non-convexities in production, I also allow for fixed operating costs as well as a sunk cost in creating a new firm.

I then extend the model to a multi-sector framework, with a finite number of sectors and a continuum of varieties in each sector. The existence of a weakly separable utility aggregator across sectors is enough to guarantee that the results of the one-sector economy carry through to the multi-sector one. The structure of the framework implies that the industry responses can be solved for independently for each sector. Furthermore, one need not impose any functional form on the sectoral aggregator as the sufficient statistic needed for aggregating welfare effects across sectors are given by the observed sectoral sales shares.

I also consider the implications of relaxing some of the other assumptions of the baseline version. Firstly, I show that generalising the 'love-of-variety' parameter introduces a wedge between the demand index that the firm cares about when setting its price and the 'true' index as given by the inverse of the marginal utility of income. In other words, the atomistic firm does not take into account the household's love-of-variety, hence the equilibrium is unaffected by this parameter. Nonetheless, this will be reflected in the welfare measure as a higher love-of-variety implies that households prefer to consumer more varieties and will therefore put a higher weight on firm entry. Mathematically, this is reflected in that the average consumer surplus measure $\mathcal{M}$ is now multiplied by the love-of-variety parameter, which in the baseline version is fixed at one.

Secondly, I consider the equilibrium and welfare implications of endogenising the household labour choice. I find that the industry equilibrium as summarised by firm's relative prices (quantities), selection and the industry demand index is unaffected by total hours worked and hence can be solved for independently. Changes in total labour supply are only reflected in the mass of firms. The industry equilibrium will nonetheless

[2] To my knowledge, no one has ever considered a demand schedule whose elasticity changes in a non-monotonic way, however this possibility is likewise covered with the sufficient statistic approach.

affect the labour decision of the household given that it determines the marginal utility of income. Specifically, a higher consumer surplus will lead to more hours worked and hence more varieties consumed in equilibrium. In terms of the welfare incidence of a shock, endogenous labour introduces an extra term to the welfare formula. Intuitively, markups introduces a wedge in the labour-leisure choice of the agent and so any shock that leads to an increase in hours worked in an economy with positive markups has a first-order welfare effect. The magnitude of this channel increases in the size of the initial wedge as measured by $\mathcal{M}$.

Lastly, I consider an extension where the production of final good varieties requires not only labour but also materials, which are sold at a price above their true labour cost. I show that introducing such a distortion in the input market is akin to the love-of-variety generalisation, showing up as a downward adjustment to the consumer surplus measure $\mathcal{M}$. In addition, the material-labour input wedge coupled with fixed labour in production produces a further welfare change on the extensive margin as changes in selection lead to changes in the aggregate labour share.

Overall, these extensions go some way to show how the model can be relaxed to allow for more realistic setups and still retain the simplicity and intuition behind the baseline decomposition of welfare.

Related Literature

The misallocation literature started with the seminal paper of Harberger (1954), while recent work includes some influential papers like Restuccia & Rogerson (2008) and Hsieh & Klenow (2009). In both of these papers however, misallocation occurs due to wedges that are either unexplained or a result of distortionary government policies. The standard CES assumption has to a large extent *hidden* the issue of misallocation in macro models because as shows by Dhingra & Morrow (2019), the constant elasticity case is the only parametrization of utility where the market outcome coincides with the social planner's solution. This work is a recent example of an important literature that generalises demand structures so as to allow for variable elasticity of substitution (Vives (1999), Feenstra

(2003), Zhelobodko et al. (2012), Weyl & Fabinger (2013)). While these papers provide important theoretical insights, they often rely on imposing further restrictions on utility or analysing cases of identical firms. Results on the direction and size of misallocation with firm heterogeneity are in general not available in the literature without relying on further constraints.

One exception is Arkolakis et al. (2019), who derive a sufficient statistic formula to quantify gains from trade. While both their formula and mine are valid in the class of monopolistic models with heterogenous firms and varying markups, they remain distinct and do not nest each-other. The reason for this is that Arkolakis et al. (2019) make two additional assumptions: the existence of a choke price above which demand for the good is zero and a Pareto distribution of firm productivities. On the other hand, they can handle a multi-country trade model and solve for the welfare incidence of trade cost shocks. I do not pursue an extension to trade in my framework, however I do allow for shocks that are much more general and which can be entirely firm-specific. [3] In spite of these important differences, both papers highlight the role that higher-order demand elasticities [4] play in welfare terms and demonstrate that one can do away with functional form assumptions and instead focus on directly recovering the firm-level sufficient statistics from the data.

Finally, this chapter is also related to the sufficient statistic in Baqaee & Farhi (2020), with the crucial distinction that in their framework markups are treated as exogenous wedges. The benefit of that assumption is that it allows them to consider a general input-output structure of the economy and derive a closed-form statistic for the distance to the Pareto frontier. However, this means that one is effectively discarding all the information content in firm markups which renders their framework not suited for policy counterfactuals.

[3] Since the sufficient statistic is derived from a first-order approximation to the equilibrium response, it holds exactly only in the limit as the vector of firm-level shocks goes to zero. The familiar caveat follows that quantifying this formula is less accurate the larger the size of the shock one wants to study.

[4] Arkolakis et al. (2019) refer to the elasticity of markup while I convexity of demand that determines output responsiveness in my work.

2.2 Framework

This section lays out the baseline version of the model. For clarity of exposition I will first present and derive the sufficient statistic in a one-sector model of the economy. This will highlight the firm-level objects one needs to recover from the data. Subsection 2.4.1 presents the multi-sector version of the model and shows that aggregating across sectors is still tractable in this framework. I also then consider how the welfare incidence formula changes when one relaxes some of the baseline assumptions. In particular, I examine extensions of the model that allow for general love-of-variety, endogenous labour choice and materials used in production.

2.2.1 Initial Equilibrium

Consumers

There is a unit mass of households who derive utility from consuming a differentiated final good, supply their unit labour inelastically and own the firms. As in Zhelobodko et al. (2012), preferences are symmetric and additively separable across varieties. Let $i \in [0, M]$ be the set of varieties available in equilibrium and p_i be their respective price. Given some total expenditure level E, the consumer chooses the optimal quantities x_i that maximise their total utility as:

$$\max_{[x_i] \geq 0} \int_0^M u(x_i)\, di \qquad \textit{subject to} \qquad \int p_i x_i\, di \leq E, \tag{2.1}$$

where $u(\cdot)$ is a three-times continuously differentiable function, strictly increasing, strictly concave and with $u(0) = 0$. Under CES preferences we would have that $u(x) = x^{\rho}$.[5] Let the wage be the numeraire so that the total expenditure of the household is given by $E = 1$.[6] The first-order

[5] Adding curvature around the linear utility aggregator is standard when using CES. Benassy (1996) illustrates how treating that curvature parameter as a free variable offers a simple way to disentangle taste-for-variety from market power which is governed by the elasticity of substitution parameter (ρ).

[6] In general, the household's total income will also be made up of profits. Because I impose a free entry condition the private sector ex-ante will make zero profits although all operating firms have positive ex-post profits.

condition to the consumer's problem gives the inverse demand function $p_i = \lambda u'(x_i)$ where λ is equal to

$$\lambda = \left(\int_0^M u'(x_i) x_i \, di \right)^{-1}. \tag{2.2}$$

Here, λ^{-1} is the Lagrange multiplier on the budget constraint and is therefore equal to the marginal utility of income. The first-order condition shows that its inverse can be re-interpreted as a demand index.

Firms

Firms produce a single variety each and are heterogenous in their variable costs. In particular, let c denote the cost type of the firm so that the amount of labour needed to produce x units of output is given by $cv(x)$. This corresponds to assuming the existence of a common production function with Hicks-neutral productivity differences across firms. I also allow for overhead costs f that are the same for all firms. The profit-maximisation problem of the firm is

$$\max_x \quad \lambda u'(x)x - cv(x) - f, \tag{2.3}$$

where I have used that $p(x) = \lambda u'(x)$. Because the firm is atomistic relative to the market, it treats the demand index λ as a constant. The presence of a positive fixed cost implies that generally some firms will be too unproductive to survive so that equilibrium features *selection.* Let c_d denote the cut-off cost level such that firms with $c > c_d$ will choose not to produce. Firm profits are decreasing in cost type and by continuity of the profit function, it must be that if equilibrium features selection then the profits of type c_d are exactly zero.

Free Entry

There is an unbounded mass of potential entrants and the type of firms is drawn from an exogenous distribution $G(c)$. An amount f_e of labour must be employed for a new firm to be created. Upon formation, the firm learns its type c and then solves the profit maximization problem given in equation (2.3). These assumptions are the same as in Hopenhayn (1992).

Free entry implies that expected profits are equal to the sunk entry cost f_e.

Market Equilibrium

Let M_e denote the mass of entering firms and from now I will denote varieties by their cost-type c. Given a distribution of types $G(c)$, fixed operating costs f and entry cost f_e, the market equilibrium is a schedule of output supply $\{x(c)\}_{c \geq c_d}$, a cost cut-off c_d, a demand index λ and an entry mass M_e such that consumers and firms behave optimally, the products and labour market all clear and this is consistent with firms having zero expected profits. The equilibrium conditions are gathered in equations (2.30) to (2.33).

$$\textit{Profit Maximisation:} \quad \lambda[u''(x(c))x(c) + u'(x(c))] = cv'(x), \tag{2.4}$$

$$\textit{Cut-off Condition:} \quad \lambda[u'(x(c_d))x(c_d)] = c_d v'(x(c_d)) + f, \tag{2.5}$$

$$\textit{Free Entry:} \quad \int_0^{c_d} \lambda u'(x(c))x(c) - cv(x(c)) - f \; dG(c) = f_e, \tag{2.6}$$

$$\textit{Resource Constraint:} \quad M_e \left(\int_0^{c_d} [cv(x(c)) + f] \; dG(c) + f_e \right) = 1. \tag{2.7}$$

Discussion. Some important theoretical properties of this framework have been studied in previous work. In particular, Dhingra & Morrow (2019) prove that a necessary condition for the market equilibrium to coincide with the first best allocation is that u is CES. In all other cases, markups will vary with firm size (type) and the decentralized market economy will be inside the Pareto frontier.[7] This makes this framework a natural environment to study misallocation in general equilibrium.

[7] They also make some interesting theoretical points in terms of the supply, selection and entry bias when the $u(\cdot)$ function satisfies certain properties but otherwise the amount of welfare losses cannot be quantified without specifying $u(\cdot)$ and the distribution of firm types.

2.2.2 Elasticities

I will now define the parameters that determine the economy's adjustment to a general perturbation in the cost distribution, as well as how these equilibrium responses map into the aggregate utility change.

Demand Side Elasticities

Let $\epsilon(x)$ and $\rho(x)$ denote the elasticity of marginal utility and the elasticity of the slope of marginal utility given by

$$\epsilon(x) \equiv -\frac{u'(x)}{xu''(x)}, \quad \text{and} \quad \rho(x) \equiv -\frac{xu'''(x)}{u''(x)}, \tag{2.8}$$

which I will refer to as elasticity and convexity respectively.[8] The demand elasticity $\epsilon(x)$ has been a prominent object in the empirical literature and it maps into the gross markup of the firm as $\mu_f = \frac{\epsilon}{\epsilon-1}$. The convexity parameter $\rho(x)$ is usually not estimated in its own right although it plays a critical role in determining the firms response to a cost or demand shock. Specifically, elasticity and convexity jointly determine the elasticity of the *marginal revenue curve*. Using the definition of firm sales, marginal revenue is given by $\lambda(u'(x) + xu''(x))$. Taking the derivative of this expression with respect to output and re-arranging, one can show that

$$\epsilon_{mr} \equiv -\frac{d\,ln(mr)}{d\,ln(x)} = \frac{2-\rho}{\epsilon-1}.$$

Supply Side Elasticities

Using the definition of variable costs as $cv(x)$, it follows that the marginal cost of a firm is equal to $cv'(x)$. Let $\epsilon_{vc}(x)$ and $\epsilon_{mc}(x)$ be the elasticity of total variable costs and the elasticity of marginal costs, respectively given by

$$\epsilon_{vc}(x) \equiv \frac{xv'(x)}{v(x)}, \quad \text{and} \quad \epsilon_{mc}(x) \equiv \frac{xv''(x)}{v'(x)}. \tag{2.9}$$

Like the demand-side elasticities, these are unit-free parameters that are purely determined by the shape of the cost function $v(\cdot)$ and do not depend on the firm-specific cost-shifter c.

[8] I follow the definitions set out in Mrázová & Neary (2017) where elasticity is defined as $\epsilon(x) = -\frac{p(x)}{xp'(x)}$ while convexity is given by $\rho(x) = -\frac{xp''(x)}{p'(x)}$. These definitions extend immediately to the monopolistic demand case since $p(x) = \lambda u'(x)$. By virtue of them being elasticities, the (multiplicative) demand shifter λ will not show up.

2.2.3 Incidence of a General Shock

Having defined the above elasticities, I can now study the firms' responses to cost shocks. Later on, a change in the tax rate can be thought of as such a cost shock. Starting from the initial distribution of costs c, consider an arbitrary non-linear shock such that the new costs are given by $c + \theta\hat{c}$, where θ parametrizes the size of the shock. The Gateaux derivative of output supply in the direction $\hat{c}$ is given by

$$\hat{x}(c) = \lim_{\theta \to 0} \frac{1}{\theta}[x(c + \theta\hat{c};\, G) - x(c;\, G)],$$

where the output response of the firm of type c takes into account the general equilibrium effects induced by the fact that other firms will also endogenously respond to the shock. We correspondingly define the response in the mass of entrants $\hat{M}_e$, the cut-off cost $\hat{c}_d$, the demand index $\hat{\lambda}$ and total utility $\hat{U}$.

Output Response

The firm-level output response is the solution to the perturbed profit maximisation condition in equation (2.30), taking into account the endogenous response of the demand index $\hat{\lambda}$. Like the requirement for the initial equilibrium, firms are assumed to correctly predict the economy's response following the shock. The output change of a firm of type c is given by

$$\frac{\hat{x}(c)}{x(c)} = \Delta(x) \left(\frac{\hat{\lambda}}{\lambda} - \frac{\hat{c}}{c} \right). \tag{2.10}$$

The shock shifts the marginal cost curve of the firm directly by $\frac{\hat{c}}{c}$. It also has an equilibrium effect due to the endogenous response of the demand index which shifts the marginal revenue curve by $\frac{\hat{\lambda}}{\lambda}$. These terms are additive because at the starting point marginal revenue is equal to marginal cost and so $\left(\frac{\hat{\lambda}}{\lambda} - \frac{\hat{c}}{c}\right)$ can be thought of as the *net cost shock* to firm c.

How this net shock is transmitted to firm output is determined by the responsiveness parameter Δ which depends on the initial size of the firm and is equal to $[\epsilon_{mr}(x) + \epsilon_{mc}(x)]^{-1}$. Since the optimal output choice is pinned down by the intersection of the marginal revenue and

marginal cost curve, the elasticities of both curves will affect responsiveness. Specifically, when either of these curves is steeper at the initial point, that will diminish the firm's responsiveness to a given shock. In the standard case of constant returns to scale and CES demand, the marginal revenue and marginal cost elasticities are constant across firms and so is the responsiveness parameter.

Selection Response

Let $\{x_d, \mu_d\}$ be the output level and markup of the cut-off firm which has a cost level of c_d. One can solve for the change in the selection margin by perturbing the zero profit condition given in equation (2.31):

$$\frac{\hat{c}_d}{c_d} = \mu_d \epsilon_{vc}(x_d) \frac{\hat{\lambda}}{\lambda}. \tag{2.11}$$

Because the first two terms in equations (2.11) are strictly positive, the sign of the selection response is solely determined by the demand index change $\frac{\hat{\lambda}}{\lambda}$. In particular, selection becomes weaker when the demand index increases and vice versa. The magnitude of the selection response also scales in the markup and variable cost elasticity of the marginal firm. Since the least productive firm's markup goes into covering the fixed cost, a larger markup indicates that a larger share of total costs are made up by the overhead component and so selection is more sensitive to changes in the demand index.

To understand the supply-side effect that shows up through ϵ_{vc}, consider the case where $\frac{\hat{\lambda}}{\lambda}$ is positive so that more firms can survive in equilibrium. Since a higher demand index implies a proportional increase in prices, the new marginal firm will be selling less output. If there are decreasing returns to scale so that $\epsilon_{vc} > 1$, the fall in output induces cost savings, pushing the profit function up and thus loosening selection further.

Demand Index Response

I obtain the endogenous response of the demand index from the first-order perturbation of the free-entry equation in (2.32). Because this formula pins down average profits in the economy, the perturbed version will

feature the firm-level output responses $\hat{x}(c)$ and the selection response $\hat{c}_d$ making it potentially intractable. However, an envelope condition implies that firm-level output changes have no first-order effect on profits and therefore will not show up in the formula for $\hat{\lambda}$.[9] Similarly, the adjustment in the selection channel will also have no first-order effect as the marginal firm makes zero profits. Therefore, the expression for $\hat{\lambda}$ depends only on the distribution of firms in the initial equilibrium and the shock itself and is equal to

$$\frac{\hat{\lambda}}{\lambda} = \frac{\text{Aggregate Variable Costs}}{\text{Aggregate Sales}} \times \int_0^{c_d} \tilde{v}(c)\frac{\hat{c}}{c}\, dc, \tag{2.12}$$

where $\tilde{v}(c) = \frac{cv(x(c))g(c)}{\int_0^{c_d} cv(x(c))g(c)dc}$ is the input share of the firm of type c. In words, the equilibrium response of the demand index is the average cost-weighted firm-level shock adjusted by the share of variable costs to total sales. The aggregate cost share adjustment is simply telling us that the competitive forces in the economy will mean that when the aggregate markup is low, the demand index will respond more to any given shock. Intuitively, when variable costs make up a larger share of sales, the shock will be attenuated to a greater extend because there is less leeway for firms to absorb it by cutting their markups. In the case where all firms sell at marginal cost and they all get the same cost shock $\frac{\hat{c}}{c} = \theta$, the demand index will be exactly equal to θ and from equation (2.10) we can see that the output produced by each firm would remain unchanged.

Mass of Entrants Response

The change in the mass of entrants is derived from the resource constraint (2.33) which is essentially a labour market clearing condition. Specifically, if more labour is used for production then fewer firms will be created in equilibrium. In Appendix A.1.3, I show that the entry response is given by

$$\frac{\hat{M}_e}{M_e} = -M_e \left(\hat{c}_d(c_d v(x_d) + f)g(c_d) + \int_0^{c_d} cv(x) \left(\frac{\hat{c}}{c} + \epsilon_{vc}(x)\frac{\hat{x}(c)}{x(c)} \right) dG(c) \right). \tag{2.13}$$

[9] In the terminology of Baqaee & Farhi (2020) we would refer to these as micro-envelope conditions. In their paper, since markups are exogenous but there are input-output linkages across firms and they have to choose how to source their inputs, micro-envelopes result from cost minimization.

The two terms in equation (2.13) correspond to the extensive (selection) and intensive margin respectively. The later one is composed of two channels: the direct effect of the cost-push shock, which requires more labour to produce the same initial levels of output, and a reallocation channel as firms optimally choose to adjust their output levels.

2.3 Welfare

Having solved for the economy's response following a general cost shock, one can use these results to get a first-order approximation of the change in welfare. Total aggregate utility at the initial equilibrium is given by

$$U = M_e \int_0^{c_d} u(x(c))\, dG(c).$$

Let u denote the average utility produced by firms in equilibrium so that the above expression can be rewritten as $U = M_e u$. The incidence on welfare is

$$\hat{U} = M_e \hat{u} + \hat{M}_e u. \tag{2.14}$$

Aggregate utility changes both because the shock induces adjustments in the production patterns $\{\hat{x}, \hat{c}_d\}$ that lead to a change in average utility per variety $\hat{u}$ and also as a result of the endogenous response in the number of varieties available as entry adjusts. To convert the utility change in money metric terms multiply $\hat{U}$ by the demand index.[10] $\lambda\hat{U}$ is the welfare measure that I will use in the rest of the section. It gives the percentage change in income required to keep the utility of the household unchanged at initial prices following the $\hat{c}$ shock.

Before showing what equation (2.14) evaluates to, let me build intuition by first discussing what happens when we fix the mass of entrants. One could think of this either as substituting the assumption of inelastically supplied labour with one of fixed entry or as representing the short-term welfare effects if the mass of entrants adjusts slowly over time.

[10]Remember that the marginal utility of income at the initial equilibrium is given by $1/\lambda$. To convert a utility change $\hat{U}$ in monetary terms, use the fact that $\Delta Income \times \text{MU}_{Income} \approx \Delta U$.

2.3.1 Fixed Entry

As previously discussed, fixing the mass of entrants does not affect the industry equilibrium because $\{c_d, x(c), \lambda\}$ are determined independently of entry. Shutting down entry and the extensive margin response, I obtain the following expression for the welfare effect

$$\lambda \hat{U} = \int_0^{c_d} \tilde{s}(c) \frac{\hat{x}(c)}{x(c)} \, dc = \int_0^{c_d} \tilde{s}(c) \Delta(c) \left(\frac{\hat{\lambda}}{\lambda} - \frac{\hat{c}}{c} \right) \, dc, \tag{2.15}$$

where $\tilde{s}(c) = \frac{s(c)g(c)}{\int_0^{c_d} s(c)g(c)dc}$ is the sales share of firms of type c. This means that the welfare impact is given by the sales-weighted output response of each firm. It also highlights that if we want to study any particular perturbation $\hat{c}$, we can get a first-order approximation as long as we have a way to recover the firm-specific output responsiveness $\Delta(c)$.

2.3.2 Average Consumer Surplus

With fixed entry, output changes at the firm level are weighted by the marginal utility from consuming that variety. Each variety's marginal utility is proportional to its price, thus giving rise to equation (2.15). In the full model with entry, one needs to weigh the firm-level output adjustment not by its own marginal utility but by how that compares to reallocating resources to creating more varieties. In other words, what will matter is the difference between the firm markup and the economy-wide or aggregate 'markup'.

In the literature so far, aggregate markup has been measured as either the sales-weighted (De Loecker et al. (2020)) or the cost-weighted (Edmond et al. (2018)) average of firm-level markups. It turns out however, that neither of these measures is what matters for weighting the benefits of reallocation. Instead, we have an equilibrium object that resembles a consumer surplus measure.

Let $\mathcal{M}$ denote the measure of *average surplus* that shows up in our welfare analysis and is given by

$$\mathcal{M} \equiv \lambda U = \frac{\int_0^{c_d} u(x(c)) \, dG(c)}{\int_0^{c_d} u'(x(c))x(c) \, dG(c)}. \tag{2.16}$$

Another way to re-express it so as to illuminate the distinction from what

has been used in the literature so far is

$$\mathcal{M} = 1 + \int_0^{c_d} \left(\frac{u(x) - u'(x)x}{u'(x)x} \right) \tilde{s}(c)\, dc,$$

where the variable being weighted can be thought of as the share of consumer surplus to the expenditure on that variety.[11] Note that it is not exactly so because actual expenditures are multiplied by the demand index λ, which however does not matter from a welfare perspective. Given the strict concavity assumption on $u(\cdot)$, the average surplus is always strictly larger than 1.

2.3.3 Welfare Decomposition with Entry

Given our solution for $\{\hat{M}_e, \hat{c}_d, \hat{x}(c), \hat{\lambda}\}$ and the definition of $\mathcal{M}$, we can decompose the total welfare effect of the cost-perturbation $\hat{c}$ as

$$\lambda \hat{U} = \overbrace{-\mathcal{M}\, M_e \int_0^{c_d} cv(x) \frac{\hat{c}}{c} dG(c)}^{direct\ effect} + \overbrace{M_e \hat{c}_d g(c_d) s_d \left[\frac{u(x_d)}{u'(x_d)x_d} - \mathcal{M} \right]}^{selection} + \overbrace{M_e \int_0^{c_d} \left[1 - \frac{\mathcal{M}}{\mu_f(c)} \right] s(c) \frac{\hat{x}(c)}{x(c)} dG(c)}^{reallocation}. \tag{2.17}$$

To fix ideas, consider a general cost-push shock so that $\hat{c}$ is positive for all firms. The first term in (2.34) is the direct effect of the shock and can be re-written as $\mathcal{M} \times \frac{\hat{\lambda}}{\lambda}$. The direct effect of a cost-push shock must always be negative as more labour is needed to produce the initial level of output which mechanically leads to a fall in the mass of varieties available. The second term is the selection effect which scales with the response in selection but also with the sales share of the cut-off firm.[12] The sign of this effect is determined by whether the utility per revenue generated by the cut-off firm is larger than or smaller than the average in the economy as given by $\mathcal{M}$. This sign is ambiguous without further restrictions on

[11] This is very similar to what Dhingra & Morrow (2019) denote as the 'social markup' and which is equal to $\frac{u(x)-xu'(x)}{u(x)}$. The only difference from the expression that appears in $\mathcal{M}$ is that the denominator is not sales but utility. They show that how social markups change relative to private markups as we increase output of a variety will be fundamental in determining the patterns of misallocation.

[12] Because total sales are equal to 1 we have that $M_e s_d g(c_d) = \frac{M_e s_d g(c_d)}{M_e \int_0^{c_d} s(c)\, dG(c)} = \tilde{s}_d$.

the utility function $u(\cdot)$. Finally, the last term gives the reallocation channel which arises due to firms adjusting their output following the shock. The expansion of a firms output will have a positive welfare effect if and only if the markup of that firm is larger than the average markup $\mathcal{M}$. Likewise the extensive channel, these inframarginal welfare effects also scale with the initial sales of the firm.

2.3.4 A Special Case: CES Demand

With CES demand, equation (2.34) reduces to the direct effect only, with the selection and reallocation channels being exactly zero. CES is the only parametrization of utility with no dispersion in firm markups even when firms are heterogenous in costs. Let ρ be the elasticity of substitution across varieties so that $\mu_f(c) = \frac{1}{\rho}$ for all firms. Using equation (2.16) one can show that whatever the underlying distribution of firms, aggregate markup is also equal to firm-level markup. This follows from the property that $\frac{u(x)}{u'(x)x}$ does not vary with output x when utility is CES. As a result, the firm weights in the reallocation channel given by $\left(1 - \frac{\mathcal{M}}{\mu_f(c)}\right)$ are zero. Similarly, the welfare weight on the selection response given by $\left(\frac{u(x_d)}{u'(x_d)x_d} - \mathcal{M}\right)$ cancels out.

The CES benchmark also illuminates another interpretation of the welfare decomposition. In particular, the direct channel in equation (2.34) can be viewed as the welfare effect from the *shift of the Pareto frontier* while the selection and reallocation channels are due to the economy moving *inside the frontier*. Because CES is the only case for which the economy is on the Pareto frontier, any other utility function will in general feature both direct and allocative welfare changes.

2.4 Extensions

This section provides extensions to the baseline version of the monopolistic model. I will in turn consider how the welfare formula changes in a multi-sector economy, generalized love-of-variety, endogenous labour and an an input wedge with material in production. Allowing for all these channels simultaneously follows straightforwardly from the individual re-

sults.

2.4.1 Multi-Sector Economy

In the benchmark case, I have focused on a single sector economy with symmetric demand. While this can be considered a good assumption for firms that produce varieties of the same good, we do not expect demand for say furniture to display the same patterns of substitution as demand for restaurants. Furthermore, firms that produce furniture are likely to be structurally different in terms of their cost structure from restaurants. As a result, it it important to extend the previous results to allow for a multi-sector model before we take it to the data.

Definitions and Aggregation

The economy is comprised of a finite number of sectors indexed by j, and a continuum of varieties within each sector indexed by the cost type c^j. There is a sector-specific utility function $u^j(\cdot)$ that determines the inverse demand function for each sector. I allow for the cost structure to be sector specific and given by $\{v^j(\cdot), f^j, f_e^j\}$. Let $U^j = M_e^j \int_0^{c_d^j} u^j(x^j(c))\, dG^j(c)$ be the total utility derived from consuming the available varieties of sector j. I will assume that households have weakly separable preferences across sectors so that the household's maximization problem is given by

$$\underset{[x_i^j]_{i\in I}}{max} \quad \mathcal{F}(U^1, U^1, \dots U^k) \qquad \textit{subject to} \qquad \sum_{j=1}^{k} M_e^j \int p^j(c) x^j(c)\, dG^j(c) \leq 1. \tag{2.18}$$

Note that we can re-write this optimisation as a two-stage problem where in the first stage the household decides the expenditures shares for each sector $\{\alpha^1, \alpha^2, \dots \alpha^k\}$ while in the second they choose the optimal bundle of varieties to consume $[x^j(c)]$ given prices and the sector-specific expenditure α^j.[13]

The market equilibrium is given by $\{M_e^j, c_d^j, x^j(c), \lambda^j\}$. Let $\{s^j, u^j\}$ respectively be the average sale and the average utility generated by firms in sector j in equilibrium. Note that consistency of budget shares

[13]The second-stage problem is therefore made up of k *independent* maximisation problems, one for each sector j.

requires that $\alpha^j = M_e^j s^j$ while the definition of aggregate sectoral utility implies that $U^j = M_e^j u^j$. The first-stage problem of the agent's utility maximization can therefore be written as

$$\max_{\{\alpha^1,\alpha^2,...,\alpha^k\}} \mathcal{F}\left(\alpha^1\frac{u^1}{s^1}, \alpha^2\frac{u^2}{s^2}, \ldots, \alpha^k\frac{u^k}{s^k}\right) \quad \text{st} \quad \sum_j \alpha^j = 1, \tag{2.19}$$

where the FOC requires that

$$\mathcal{F}'_j\frac{u^j}{s^j} - \frac{1}{\psi} = 0. \tag{2.20}$$

The utility impact of any shock to the first order is given by

$$\hat{U} = \sum_{j=1}^{k}\left(\mathcal{F}'_j\frac{u^j\alpha^j}{s^j}\right)\left[\frac{\hat{\alpha}^j}{\alpha^j} + \frac{\hat{u}^j}{u^j} - \frac{\hat{s}^j}{s^j}\right].$$

Multiplying both sides of the equation by the inverse of the Lagrange multiplier of the budget constraint and using the optimality condition for consumption shares we get that

$$\psi\hat{U} = \sum_{j=1}^{k}\alpha^j\left[\frac{\hat{\alpha}^j}{\alpha^j} + \frac{\hat{u}^j}{u^j} - \frac{\hat{s}^j}{s^j}\right]. \tag{2.21}$$

Let us now discuss the implications of this result. Firstly, note that by construction $\sum_{j=1}^{k}\hat{\alpha}^j = 0$ [14] which implies that the first term will disappear from the welfare statistic. Intuitively, the re-allocation of consumption across sectors does not have a first-order effect on aggregate utility because consumption shares are chosen optimally with the marginal utility of spending one more pound equalised across sectors. Secondly, we can show that the term $\frac{\hat{u}^j}{u^j} - \frac{\hat{s}^j}{s^j}$ is simply equal to the welfare metric in equation (2.34) divided by the sector specific aggregate markup $\mathcal{M}^j$. In other words, the multi-sector economy behaves like k different one-sector economies where the sufficient statistic for aggregating across sectors is given by the observed expenditure shares.

2.4.2 Generalized Love-of-Variety

As Benassy (1996) has pointed out, the standard monopolistic model imposes a very tight link between the 'love-of-variety' parameter and the

[14] This result is no longer true when labour supply is endogenous and so the total amount of hours adjusts following a shock. I discuss the implications of relaxing this assumption in subsection 2.4.3.

degree of market power (elasticity of substitution) that has important implications on the conclusions one draws from these models. For example, Montagna (2001) shows that once you disentangle the two, the parameter governing 'love-of-variety' can determine whether trade has positive or negative welfare effects depending on whether the efficiency or the variety effect dominate. Following the definition in Benassy (1996), let 'love-of-variety' be the ratio of the utility derived from producing a number of goods M in equal amounts relative to producing a single good and given by $v(M) = \frac{Mu(L/M)}{u(L)}$. In the setup defined above, we have that the elasticity with respect to the number of varieties is given by

$$\frac{Mv'(M)}{v(M)} = 1 - \epsilon_u \left(\frac{L}{M}\right). \tag{2.22}$$

Unlike the CES case where the elasticity ϵ_u is independent of the quantity consumed, the generalized monopolistic model will feature varying 'love-of-variety'. Nonetheless, one might still want to parametrize this later quantity independently from the elasticity of substitution function that like ϵ_u also depends on the parametrization of the utility function. A simple way to model love-of-variety in a more flexible way would be to add some curvature around the total mass of varieties available to the consumer. In particular, let agents care about the total mass of varieties that they consume in equilibrium according to some function $H(\cdot)$ so that the utility aggregator is now given by

$$U = H(M_e) \int_0^{c_d} u(x(c))\, dG(c)$$

where in the benchmark case we had that $H(M_e) = M_e$. [15] Let the Lagrange multiplier for the budget constraint be $\tilde{\lambda}$ with the optimality condition now equal to $p(x) = \frac{H(M_e)}{M_e}\tilde{\lambda}u'(x)$. Multiplying both sides by the firm quantity x and integrating over all varieties, the expression for the marginal utility of income becomes

$$\tilde{\lambda} = \left(H(M) \int_0^{c_d} u'(x(c))x(c)\, dG(c)\right)^{-1}. \tag{2.23}$$

[15] Note that one can always re-write the integral over the mass of varieties in equation (2.1) as an integral over firm types c as identical firms will charge the same price in equilibrium and therefore the household will consume the same quantity $x(c)$.

Again, if one chooses $H(M_e) = 1$ (or any constant function) we are back to the baseline expression. The inverse demand function is given by

$$p(x) = \frac{H(M_e)}{M_e}\tilde{\lambda}u'(x) = \frac{u'(x)}{M_e \int_0^{c_d} u'(x(c))x(c)\, dG(c)}. \tag{2.24}$$

Note that the inverse demand function takes the same form as in the benchmark model and in particular, the function $H(\cdot)$ does not appear in this expression. This implies that we can define the demand index similarly to before as $\lambda = \left(M_e \int_0^{c_d} u'(x(c))x(c)\, dG(c)\right)^{-1}$.

It also means that there is now a wedge between the price index that matters to the firm $1/\lambda$ and the marginal utility of income $1/\tilde{\lambda}$, with $H(M_e)\tilde{\lambda} = M_e\lambda$. We conclude that given a utility function $\{u(\cdot)\}$, a cost function $\{v(\cdot)\}$, a distribution of firm-level cost shifters $G(c)$ and the fixed and entry costs $\{f, f_e\}$, varying love-of-variety has no effect on the equilibrium outcome. Nonetheless, how much households care about consuming different varieties will matter for welfare.

As before, let u denote the average utility produced by firms in equilibrium so that aggregate utility is $U = H(M_e)u$. The first-order welfare effect of any general shock is equal to $\hat{U} = H'(M_e)\hat{M}_e u + H(M_e)\hat{u}$. To convert the utility change into the standard money metric measure, divide by the marginal utility of income to get

$$\begin{aligned}\tilde{\lambda}\hat{U} &= \frac{M_e}{H(M_e)}\lambda\left(H'(M_e)\frac{\hat{M}_e}{M_e}U + H(M_e)\hat{u}\right) \\ &= \eta_e\mathcal{M}\frac{\hat{M}_e}{M_e} + M_e\lambda\hat{u}.\end{aligned}$$

where $\eta_e = \frac{M_e H'(M_e)}{H(M_e)}$ is the elasticity of the love-of-variety function. This expression is the same as the baseline equation (2.14) but aggregate markup $\mathcal{M}$ is now multiplied by the love-of-variety parameter η_e to give the true welfare weight on firm entry. The welfare decomposition given in (2.34) changes solely by replacing the consumer surplus measure $\mathcal{M}$ by $\eta_e\mathcal{M}$ while the intuition for each channel remains unchanged.

2.4.3 Endogenous Labour Supply

Let l be the labour supplied by the representative household and $\varphi(l)$ be the additive disutility cost from working. The household's objective function is now given by

$$\underset{\{[x_i]\geq 0,l\}}{max} \quad \int_0^M u(x_i)\, di - \varphi(l) \qquad subject\ to \qquad \int p_i x_i\, di \leq l \quad (2.25)$$

where I have normalized the wage to 1 so that total household income is l. [16]

Equilibrium Conditions The first three equations (2.30) - (2.32) (profit maximization, cut-off condition, free entry) remain unchanged and together with the following two conditions determine the equilibrium

$$Resource\ Constraint: \quad M_e \left(\int_0^{c_d} [cv(x(c)) + f] dG(c) + f_e \right) = l, \tag{2.26}$$

$$Labour\text{-}Leisure\ Choice: \quad \varphi'(l) = \frac{M_e \int_0^{c_d} u'(x(c)) x(c)\, dG(c)}{l}. \tag{2.27}$$

One important feature of this model is that the industry equilibrium as characterized by $\{x(c), c_d, \lambda\}$ is independent of the mass of entrants and the labour supply and can be solve for using (2.30) - (2.32). [17]

Remember that the welfare measure in the baseline version was defined as the percentage change in income at initial prices. Since income here is endogenous and given by l, the equivalent expression is $\frac{\lambda \hat{U}}{l}$.

Consider two equilibria that are identical, but where in one of them labour is optimally supplied while in the other it is exogenously fixed at 1. The welfare effect in the endogenous labour version is

$$\frac{\lambda \hat{U}}{l} = \frac{\lambda \hat{U}^{\mathrm{l=1}}}{l} + (\mathcal{M} - 1)\frac{\hat{l}}{l}. \tag{2.28}$$

where $\hat{U}^{\mathrm{l=1}}$ is the utility change in the benchmark case. The industry equilibrium response and hence the average utility change is independent on whether and how much labour (and the mass of entrants) adjusts so we get the same first term as in the benchmark. In addition, labour changes have a further welfare effect when the economy is not on the Pareto

[16] The full expression for household income is $wl + \Pi$. However, the free entry condition implies that the corporate sector will exhaust its profits to cover the fixed entry so that $\Pi = 0$.

[17] This would not in general be the case if the utility of consumption and the disutility of work were not separable.

Frontier. As $\mathcal{M}$ represents the wedge on working induced by prices being above marginal costs, households work too little in equilibrium and so an increase in hours worked generates a first-order welfare increase.

2.4.4 Materials in Production

Consider now the case where in order to produce final good varieties firms need materials as well as labour. I will assume that materials are produced with a linear technology using labour only and are sold to final good producers at a price p_m that can be strictly greater than the labour cost given by 1. [18]

With material inputs, the firm's problem now becomes a two-stage one, where in the first stage the firm chooses inputs to minimise costs and in the second stage the firm decides how much output to produce to maximize its profits. With Hicks-neutral productivity differences and a production function that is homogenous in labour and materials (more details in Appendix A.1.5), we recover a cost function that is very similar to the benchmark case

$$vc = \psi(p_m)cv(x). \tag{2.29}$$

The price of material inputs p_m will of course matter for determining variable costs, having the nice feature that it enters multiplicatively through the function $\psi(\cdot)$.[19]

Let η^* be the labour to material ratio that firms optimally choose in equilibrium. Because of the homogeneity assumption, this ratio will be the same for all firms and will depend on p_m only. The equilibrium

[18] The production function of material goods can be generalized, however that would not add much to the intuition of this mechanism. Also note that profits from these firms are redistributed to the representative household but do not affect the equilibrium outcome when labour is exogenously fixed.

[19] On the other hand, one could assume a particular functional form for production and solve explicitly for the cost function.

conditions are now given by

$$\textit{Profit Maximisation:} \quad \lambda[u''(x(c))x(c) + u'(x(c))] = \psi(p_m)cv'(x), \tag{2.30}$$

$$\textit{Cut-off Condition:} \quad \lambda[u'(x(c_d))x(c_d)] = \psi(p_m)c_d v'(x(c_d)) + f, \tag{2.31}$$

$$\textit{Free Entry:} \quad \int_0^{c_d} \lambda u'(x(c))x(c) - \psi(p_m)cv(x(c)) - f \; dG(c) = f_e, \tag{2.32}$$

$$\textit{Resource Constraint:} \quad M_e \left(\int_0^{c_d} \left[\frac{1+\eta^*}{p_m+\eta^*} \psi(p_m)cv(x(c)) + f \right] \; dG(c) + f_e \right) = 1. \tag{2.33}$$

Note that if the price of the material inputs was not distorted so that $p_m = 1$, we could interpret this more general model as one where labour is the only variable input, and the firm-specific cost shifters are given by $\tilde{c} = \psi(1)c$. This is because the new equilibrium conditions are such that the variable and marginal costs of all firms are now multiplied by $\psi(p_m)$, with the exception of the resource constraint that also features an adjustment factor $\frac{1+\eta^*}{p_m+\eta^*}$. This implies the economy's response to the same proportional shock $\left(\frac{\hat{c}}{c}\right)$ gives rise to identical responses and welfare change.

When materials are priced at a premium, the resource constraint is not equivalent any longer and the adjustment in the mass of entrants will differ from the benchmark. [20] This gives rise to the following welfare change

$$\begin{aligned}\lambda \hat{U} = &- \mathfrak{q}\mathcal{M} \, M_e \int_0^{c_d} cv(x) \frac{\hat{c}}{c} dG(c) + M_e \hat{c}_d g(c_d) s_d \left[\frac{u(x_d)}{u'(x_d)x_d} - \mathfrak{q}\mathcal{M} + (1-\mathfrak{q})\mathcal{M} \frac{f}{s_d} \right] \\ &+ M_e \int_0^{c_d} \left[1 - \frac{\mathfrak{q}\mathcal{M}}{\mu_f(c)} \right] s(c) \frac{\hat{x}(c)}{x(c)} dG(c),\end{aligned} \tag{2.34}$$

where $\mathfrak{q} = \frac{1+\eta^*}{p_m+\eta^*}$ is the distortion introduced by input markups, which disappear whenever materials are priced at marginal cost. This is the

[20] Another potential margin of adjustment is to model the input price response following a general shock rather than assume it is 0, however that is beyond the scope of this paper.

same formula as equation (2.34) in the baseline version, with an adjustment for the average surplus $\mathcal{M}$ by $\mathfrak{q}$ and an extra term in the selection effect. The presence of input markups leads to a downward adjustment for the average surplus provided by final good producers as the markup represents a wedge to production efficiency. For a given input mix, the downward adjustment will be smaller for lower input markups, while for a given price p_m, a higher labour to material share η^* attenuates the adjustment. [21]

The extra term in the selection channel shows up due to the existence of overhead labour costs in production, which are not subject to a wedge. Therefore, a loosening of selection produces a positive effect through the implicit higher apportion to labour inputs due to the overheads. As expected, this effect will be larger whenever overhead labour accounts for a larger share of total production costs $\frac{f}{s_d}$ and for a higher wedge (i.e $\mathfrak{q}$ being further away from 1). Absent this last channel, allowing for materials in production priced at a positive markup is equivalent to the elasticity of the love-of-variety function H being lower than 1.

2.5 Conclusion

This chapter revisits the ubiquitous monopolistic competition model and proposes a new statistic to evaluate welfare changes induced by general supply-side shocks. Building upon the results on partial equilibrium demand-side sufficient statistics in Mrázová & Neary (2017), I demonstrate that a transparent and intuitive formula can also be derived in general equilibrium, with an unrestricted distribution of firm types.

The heterogeneity of firms implies that the indirect effect of a general shock will have both an extensive and an intensive margin. The extensive margin shows up whenever there is selection in equilibrium, which in turn means that a change in 'competitiveness' following a shock, will change who is the least productive firm in equilibrium. The intensive margin relates to output responses over the whole distribution of firms and is

[21] The labour material share is in itself an endogenous object so one could instead take the total derivative $\frac{\partial \mathfrak{q}}{\partial p_m} = \frac{\frac{\partial \eta}{\partial p_m}(p_m - 1) - (1+\eta)}{p_m + \eta}$ but the effect is in general ambiguous without specifying the production function.

nicely summarized by the firm's markup and responsiveness parameter. Whether these channels contribute positively or negatively to welfare, depends crucially on the average surplus measure $\mathcal{M}$. This formula also clarifies that markups alone are not sufficient firm-level statistics for welfare analysis unless one specifies a parametric form of demand. Furthermore, the average economy-wide markup, which is often reported as a headline figure in many papers is different from the benchmark $\mathcal{M}$ that matters for welfare. [22]

Finally, I look at extensions of the baseline model in directions that I judge to be economically interesting and potentially important from an empirical point of view. The intuition of the formula remains very clear and the new channels show up as either additive terms or adjustments to the benchmark value of $\mathcal{M}$. Importantly, extending to a multi-sector model preserves the simplicity of the formula as one can solve separately for the industry responses in each sector and then aggregate the industry-level welfare effects using the observed sales shares.

[22] Of course, it could be the case that these two measure correlate highly for particular functional forms and productivity distributions but there are no conditions to guarantee so with the sole exception of CES, where they are exactly the same.

Appendix A

Appendix to Chapter 2

A.1 Model Derivations

A.1.1 Output Response

The output of a firm of type c following the general cost-perturbation $\hat{c}$ is given by $x(c)+\mu\hat{x}(c)$ and is determined by the solution to the perturbed first-order condition. For clarity of notation, I suppress the notation of output as a function of costs $x(c)$ and simply use x instead. Taking a first order approximation to $\lambda(xu''(x)+u'(x))=cv'(x)$ we get

$$[\lambda+\mu\hat{\lambda}][u'(x+\mu\hat{x})+(x+\mu\hat{x})u''(x+\mu\hat{x})]=[c+\mu\hat{c}][v'(x+\mu\hat{x})]$$

$$[\lambda+\mu\hat{\lambda}][u'(x)+xu''(x)+\mu(\hat{x}u''(x)+\hat{x}u''(x)+\hat{x}xu'''(x))]=[c+\mu\hat{c}][v'(x)+\mu\hat{x}v''(x))]$$

$$\lambda\hat{x}[2u''(x)+xu'''(x)]+\hat{\lambda}[u'(x)+xu''(x)]=c\hat{x}v''(x)+\hat{c}v'(x)$$

$$\lambda[u'(x)+xu''(x)]\left(\frac{\hat{x}}{x}\frac{x[2u''(x)+xu'''(x)]}{u'(x)+xu''(x)}+\frac{\hat{\lambda}}{\lambda}\right)=cv'(x)\left(\frac{\hat{x}}{x}\frac{xv''(x)}{v'(x)}+\frac{\hat{c}}{c}\right)$$

The first terms on both sides cancel as they simply equal the initial equilibrium condition ($MR=MC$). Using the definition of ϵ and ρ, the term multiplying the output response on the LHS is

$$\frac{x[2u''(x)+xu'''(x)]}{u'(x)+xu''(x)}=\frac{xu''(x)\left[2+\frac{xu'''(x)}{u''(x)}\right]}{xu''(x)\left[\frac{u'(x)}{xu''(x)}+1\right]}=\frac{2-\rho}{1-\epsilon}.$$

When defining the elasticity of the marginal revenue curve I will add a negative sign which together with the assumption of firm optimality implies that the ϵ_{mr} will always be positive.

$$\epsilon_{mr}=-\frac{d\,log\,(xu''+u')}{d\,log\,x}=-(u''+xu'''+u'')\frac{x}{xu''+u'}=\frac{x[2u''+xu''']}{xu''+u'}=\frac{2-\rho}{\epsilon-1}.$$

Using the definition of the marginal cost elasticity and putting it all together we have that

$$\begin{aligned}\frac{\hat{x}}{x}\left[-\epsilon_{mr}-\epsilon_{mc}\right]+\frac{\hat{\lambda}}{\lambda}&=\frac{\hat{c}}{c}\\ \frac{\hat{x}}{x}&=\left[\epsilon_{mr}+\epsilon_{mc}\right]^{-1}\left(\frac{\hat{\lambda}}{\lambda}-\frac{\hat{c}}{c}\right),\end{aligned}$$

which is exactly equation (2.10).

A.1.2 Demand Index Response

Let $\pi(\lambda, c)$ be the optimized profit function and let $\tilde{\lambda}$ denote the perturbed demand index. We derive the first order perturbation in the free entry condition as

$$\begin{aligned}\frac{\mathbb{E}[\pi(\tilde{\lambda},\tilde{c})]-\mathbb{E}[\pi(\lambda,c)]}{\mu}=&\frac{1}{\mu}\left\{\int_0^{c_d+\mu\hat{c}_d}\pi(\lambda+\mu\hat{\lambda},c+\mu\hat{c})\;dG(c)-\int_0^{c_d}\pi(\lambda,c)\;dG(c)\right\}\\ \underset{\mu\to 0}{=}&\frac{1}{\mu}\left\{\int_0^{c_d}\left[\pi(\lambda,c)+\mu\hat{\lambda}\,\pi'_\lambda+\mu\hat{c}\,\pi'_c\right]dG(c)+\mu\hat{c}_d g(c_d)\pi(\lambda,c_d)\right.\\ &\left.-\int_0^{c_d}\pi(\lambda,c)\;dG(c)\right\}\\ =&\int_0^{c_d}\left[\hat{\lambda}\,\pi'_\lambda+\hat{c}\,\pi'_c\right]dG(c),\end{aligned}$$

where we have used the fact that the cut-off firm must be making exactly zero profits so the last term in the second line cancels. On the other hand, if there is no cut-off c_d in the initial equilibrium, one does not need to consider the selection channel.

Apply the envelope theorem on the profit function to get $\{\pi'_\lambda, \pi'_c\}$ and solve for $\hat{\lambda}$ by setting the above expression to zero. If we wanted to extend the perturbation to allow for a change in the cost of entry that would be done easily by equating to $\hat{f}_e$.

$$\begin{aligned}\int_0^{c_d}\hat{\lambda}\,u'(x)x+\hat{c}\,(-v(x))\;dG(c)&=0\\ \hat{\lambda}\int_0^{c_d}u'(x)x\;dG(c)&=\int_0^{c_d}\frac{\hat{c}}{c}cv(x)\;dG(c)\\ \frac{\hat{\lambda}}{\lambda}=\frac{\int_0^{c_d}cv(x)\;dG(c)}{\lambda\int_0^{c_d}u'(x)x\;dG(c)}&\int_0^{c_d}\frac{\hat{c}}{c}\,\frac{cv(x)}{\int_0^{c_d}cv(x)\;dG(c)}\;dG(c),\end{aligned}$$

where the weights for the cost shock are just given by the variable cost weight of the firm of type c. To get the correction term given in equation

(2.12) multiply both integrals by M_e and apply the definitions of total variable costs and total sales.

Selection Response

To solve for $\hat{c}_d$ in an equilibrium with selection we again turn to the profit function and we use the fact that $\pi(\lambda + \mu\hat{\lambda}, c_d + \mu\hat{c}_d) = 0$.

$$
\begin{aligned}
&s_d\frac{\hat{\lambda}}{\lambda} + \frac{\hat{x}_d}{x_d}\left[s_d\left(1-\frac{1}{\epsilon_d}\right) - c_d v(x_d)\epsilon_c\right] - c_d v(x_d)\frac{\hat{c}_d}{c_d} = 0\\
&s_d\frac{\hat{\lambda}}{\lambda} + \frac{\hat{x}_d}{x_d}\left[s_d\left(1-\frac{1}{\epsilon_d}\right) - \frac{s_d}{\mu_d\epsilon_c}\epsilon_c\right] - \frac{s_d}{\mu_d\epsilon_c}\frac{\hat{c}_d}{c_d} = 0\\
&\frac{\hat{\lambda}}{\lambda} + \frac{\hat{x}_d}{x_d}\left[\left(1-\frac{1}{\epsilon_d}\right) - \frac{1}{\mu_d}\right] - \frac{1}{\mu_d\epsilon_c}\frac{\hat{c}_d}{c_d} = 0
\end{aligned}
$$

The envelope implies that output adjustments for the cut-off firm will not affect the firm's profit since firms are always making zero profits on the marginal unit of output that they sell and hence all the effects come from the adjustment in the demand index.

A.1.3 Mass of Entrants Response

Let's re-write the resource constraint as $M_e\vartheta = 1$ where ϑ is the average labour used by a variety in equilibrium where here variety includes also those that do not produce any good. We derive the first-order perturbation in ϑ as

$$
\begin{aligned}
\frac{\tilde{\vartheta}-\vartheta}{\mu} &\underset{\mu\to 0}{=} \frac{1}{\mu}\left\{\int_0^{c_d+\mu\hat{c}}[(c+\mu\hat{c})v(x+\mu\hat{x})+f]\,dG(c) + f_e\right.\\
&\qquad \left. - \left(\int_0^{c_d}[cv(x)+f]\,dG(c) + f_e\right)\right\}\\
&= \int_0^{c_d}[\hat{c}v(x) + \hat{x}cv'(x)]\,dG(c) + \hat{c}_d g(c_d)[c_d v(x_d)+f]\\
&= \int_0^{c_d} cv(x)\left(\frac{\hat{c}}{c} + \frac{xv'(x)}{v(x)}\frac{\hat{x}}{x}\right)\,dG(c) + \hat{c}_d g(c_d)[c_d v(x_d)+f].
\end{aligned}
$$

Rearranging and using the fact that the derivative must be zero since the total resources are fixed we get the expression for the change in the mass of firms

$$
\begin{aligned}
\frac{\hat{M}_e}{M_e} &= -M_e\left(\hat{c}_d g(c_d)[c_d v(x_d)+f] + \int_0^{c_d} cv(x)\left(\frac{\hat{c}}{c} + \epsilon_c\frac{\hat{x}}{x}\right)\,dG(c)\right)\\
&= -\left(M_e\hat{c}_d g(c_d)s_d + \frac{\hat{\lambda}}{\lambda} + M_e\int_0^{c_d} cv(x)\epsilon_c\frac{\hat{x}}{x}\,dG\right).
\end{aligned}
$$

A.1.4 Change in Utility

To derive the impact on utility we can use the mass of entrants response together with the change in average utility of a variety which is

$$\begin{aligned}\frac{\tilde{u}-u}{\mu} &= \frac{1}{\mu}\left\{\int_0^{c_d+\mu\hat{c}} u(x+\mu\hat{x})\ dG(c) - \int_0^{c_d} u(x)\ dG(c)\right\} \\ &\underset{\mu\to 0}{=} \frac{1}{\mu}\left\{\int_0^{c_d} u(x)+\mu u'(x)\hat{x}\ dG(c) + \hat{c}_d u(x_d)g(c_d) - \int_0^{c_d} u(x)\ dG(c)\right\} \\ &= \hat{c}_d u(x_d)g(c_d) + \int_0^{c_d} xu'(x)\frac{\hat{x}}{x}\ dG(c).\end{aligned}$$

We substitute these expressions in $\hat{U} = u\hat{M}_e + M_e\hat{u}$, multiply by λ to convert into monetary units and re-arrange the terms to get the welfare decomposition in equation 2.34.

A.1.5 Materials in Production

Let ω denote the productivity level of a firm and the production function be $F(l,m)$. Homogeneity of F implies that $F(\theta l, \theta m) = \theta^r F(l,m)$. The cost minimization problem of the firm is

$$\min_{l,m} \quad p_m m + l \quad \text{st} \quad \omega F(m,l) \geq x.$$

Combining the first order conditions for labour and capital, we have that

$$\begin{aligned}p_m &= \frac{F'_m(m,l)}{F'_l(m,l)} \\ &= \frac{m^{r-1}F'_m(1,l/m)}{m^{r-1}F'_l(1,l/m)} \\ &= \frac{F'_m(1,l/m)}{F'_l(1,l/m)},\end{aligned}$$

where in the second step, I use the fact that since F is homogenous of degree r, its first derivatives are homogenous of degree $r-1$. From this, we conclude that the optimal ratio of labour to materials $\eta^* = \frac{l^*}{m^*}$ only depends on the inputs relative prices and is independent of the productivity level of the firm ω or the desired output x. Use the binding constraint to solve for the optimal m^*

$$\begin{aligned}x &= \omega F(m^*, l^*) = \omega m^{*r} F(1, l^*/m^*) \\ m^* &= \left(\frac{x}{\omega F(1, \eta^*(p_m))}\right)^{1/r}.\end{aligned}$$

Finally, solve for the minimized cost function of the firm

$$\begin{aligned} VC(x) &= m^* + p_m l^* = m^*(1 + p_m \eta^*(p_m)) \\ &= \left(\frac{1 + p_m \eta^*(p_m)}{F(1, \eta^*(p_m))} \right) \omega^{-1/r} x^{1/r}. \end{aligned}$$

This is exactly the cost function given in the main text in equation (2.29), where $\psi(p_m)$ is given by the term in brackets and the cost shifter c is equal to $\omega^{-1/r}$.

Chapter 3

Correcting Market Power with Taxation: An Application to the UK

3.1 Introduction

What can tax policy do to alleviate the distortions caused by market power? Should taxes be raised for large and powerful firms or small unproductive ones? Empirical work documents substantial heterogeneity across firms, with growing evidence that the disparity has gone up in the last decades. The dispersion in markups suggests that welfare gains are possible by improving the allocative efficiency of the market. Specifically, differential taxation of sales can be used to affect firms' pricing decisions and hence move the equilibrium closer to the Pareto Frontier.

To quantify misallocation losses in general equilibrium and study policy interventions, researchers have traditionally relied on functional form assumptions on the unobserved demand schedules to discipline the data. Instead, I propose a new non-parametric estimator for firm output responsiveness which together with the standard estimator of markups can be used to evaluate a sufficient statistic formula that does not make any parametric assumption on demand. In particular, I adjust the formula derived in Chapter 2 to study the welfare effects of arbitrary tax changes. I use a long-running official survey of British businesses to estimate firm-level markup and output responsiveness. With these empirical findings

and my welfare formula at hand, I evaluate the welfare gains from a simple revenue-neutral two-tier VAT reform. I find that the distribution of firm-level welfare weights is such that it is welfare-improving to subsidise small firms at the expense of large ones and even a two-tier tax rate can deliver substantial gains to the consumer.

I use the structure provided by symmetric monopolistic competition with Hicks-neutral productivity differences to derive a novel identification of firm-level output responsiveness. In particular, assume that variable inputs display fixed returns to scale in production denoted by r. Exploiting a scalar unobservable assumption in firm-level costs one can show that variable costs (VC) and sales (S) in the cross-section of firms must satisfy

$$\frac{\partial VC}{\partial S} = 1 - r\Delta^{-1}. \tag{3.1}$$

The intuition for equation 3.1 is as following. Optimality requires that a firm makes zero profits on the marginal unit sold so as sales change, variable costs must change one-for-one and hence the constant term. However, for sales to increase at any point in time it must be that the firm is moving down the demand curve. Evidently, higher sales can be achieved by different combinations of changes in quantity and changes in price. A higher output responsiveness (Δ) means that a given increase in sales is concurrent with a larger output adjustment, hence variable costs are increasing faster. On top of that, if returns to scale are not constant ($r \neq 1$), an *endogenous* supply-side effect kicks in. Specifically, for any given output response as determined by Δ, faster decreasing returns to scale ($r \downarrow$) means that variable costs must increase by more.

I use the data from the UK's *Annual Business Survey* to estimate non-parametrically the slope of variable costs to sales at the industry level and recover Δ by inverting equation 3.1. Note that in the empirical application, I assume knowledge of the returns to scale parameter but I show in robustness exercises that the results remain qualitatively unchanged unless the returns to scale are very strongly increasing. [1] To estimate

[1] Basu & Fernald (1997) show that the estimation of the returns to scale depends non-trivially on the level of industry aggregation used as well as on whether one

markups (μ_f), I rely on the standard cost-minimisation approach pioneered by De Loecker & Warzynski (2012) although the assumption of fixed returns in variable inputs (potentially a subset of all inputs) allows one to sidestep estimating a production function, as the elasticity of costs is independent of the input mix and does not vary by firm.

The empirical results are presented and discussed in detail in Section 3.4. The main takeaways are as following. Firstly, firm markups are generally decreasing with firm size at the industry level. This is a novel result and in disagreement with the usual demand parametrizations in the literature that postulate a positive relationship. For example, Edmond et al. (2018) calibrate a Kimball demand that has a positive markup-size slope following the observed positive relationship between labour's revenue productivity and size. The use of materials in the bundle of inputs switches the sign of this relationship. I confirm this finding by estimating the superelasticity parameter under a Kimball demand assumption using either labour or the bundle of labour and materials as the variable input. In all industries, I get positive superelasticities when using labour only and negative superelasticities when using the bundle input. The fit of the non-linear regression as measured by the R^2 drops drastically when using labour only as opposed to the labour plus material bundle. Secondly, I find that output responsiveness increase with firm size. In practice this means that for small firms, the percentage difference in (unobserved) output for a given percentage difference in productivity is smaller than for larger, more productive firms. This also has the implication that following a common shock, large firms would adjust their output by more than smaller ones. Finally, one can recover price pass-through at the firm level from the markup and the output responsiveness. I find that the price pass-through to a cost shock is on average decreasing with firm size. These results are well in line with findings in the trade literature as in Amiti et al. (2019), although a direct comparison of the magnitude is not applicable.

assumes a gross output or value-added production function. However, they still conclude that most 2-digit industries in the US display slightly decreasing returns to scale and those are the returns to all inputs including capital while my assumption regards only the variable input part, i.e. labour and materials.

In Section 3.5 I lay out the tax reform application. As a starting point, I consider the (unrealistic) case where the government can impose a linear firm-specific sales tax.[2] I derive the welfare incidence formula for changing the tax rate of any single firm in isolation. The reallocation channel for this *elementary tax reform* results from the fact that a firm whose tax rate is increased chooses to supply less output and therefore uses less resources. The labour that is freed up will be employed by other firms in equilibrium such that production is shifted to the 'average' firm. The welfare effect is thus determined by the difference between the *welfare weight* of the shocked firm and the average (sales-weighted) weight in the economy. This implies that reallocation can be welfare improving if and only if there is dispersion in firm-level welfare weights which are given by

$$\omega = \left(1 - \frac{\mathcal{M}}{\mu_f}\right) \Delta, \tag{3.2}$$

where $\mathcal{M}$ is a measure of average consumer surplus in the initial equilibrium.

Although it serves as the 'average markup' against which to compare firm level markups and hence evaluate welfare gains from reallocation, this measure is not the same as the aggregate markup defined in previous work as either a cost-weighted or sales-weighted average of firm markups (Edmond et al. (2018), De Loecker et al. (2020)). Furthermore, one cannot learn about $\mathcal{M}$ from knowledge of the distribution of markups only, so in the tax application I treat it as a parameter to calibrate.

For the UK, I find that welfare weights decrease with firm size for any calibration of the unobserved average surplus $\mathcal{M}$.[3] This result holds both across industries generally and over time for the sample years of $1997 - 2010$. I use the approximate monotonicity of ω against sales

[2]Given the UK context, I study a reform of the VAT tax rather than a sales tax as implemented in the US. Firms pay VAT on the totality of their sales to the final consumer but are reimbursed for any VAT paid to their suppliers. This implies that the two taxes are different only if intermediate good producers have market power. If that market power is homogenous then we can provide a simple mapping between the two taxes by taking into account the pass-through of intermediaries in the VAT tax case. For the sake of simplicity, I will use sales and VAT taxes interchangeably.

[3]By concavity of the utility function, the average consumer surplus is bounded below by 1.

in the empirical results to restrict the tax reform to an economy-wide two-tier bracket tax change. Furthermore, I impose that the tax reform be revenue neutral. Since welfare weights are falling in firm size, it is welfare-improving to tax large firms and subsidise small ones. The *welfare multiplier* [4] of this reform will depend on the sales threshold at which the tax rate jumps. Nonetheless, I find that the multiplier is positive for a large set of thresholds and for any value of $\mathcal{M}$ larger than 1. In this sense, the proposed tax reform is robust to the choice of large versus small firms and the calibration of the unobserved average surplus.

For the benchmark case of $\mathcal{M} = 1.2$, I estimate that increasing the VAT rate from 20% to 24% for firms with sales greater than £2m and giving a tax cut to smaller firms leads to an increase in aggregate utility of around 2%. This figure increases in the average surplus $\mathcal{M}$ and is bounded below by 1.1%. Overall, my findings support a tax relief for small and medium sized firms, at the expense of higher taxes for larger ones.

Related Literature

This paper is motivated by the recent literature on increasing firm concentration and falling labour share. Methodologically, it relates to the empirical literature on estimating markup and price pass-through at the firm or product level as well as the more theoretical and quantitative work on misallocation.

A large set of papers document (Karabarbounis & Neiman (2013)) and try to explain (Karabarbounis & Neiman (2018), Rognlie (2016), Barkai (2020)) the fall in the aggregate labour share that started around 1980. Empirical studies at the firm-level document an increase in firm-level dispersion whether that is measured by market shares in Autor et al. (2020), TFP in Decker et al. (2018) or markups in De Loecker et al. (2020). The reallocation of production and rise in concentration has also been documented by Rossi-Hansberg et al. (2018) and Kehrig & Vincent (2021), while Gutiérrez & Philippon (2017) provide evidence that higher

[4] Recall that I solve the model using a first-order perturbation so the welfare effects scale linearly in the size of the shock θ, hence we can talk of a *welfare multiplier*. By definition, the equations hold exactly only in the limit as $\theta \to 0$.

concentration has lead to a fall in business investment.

The methodology of estimating markups from cost minimization has been pioneered by Hall (1988) and extended to a firm-level approach by De Loecker & Warzynski (2012). De Loecker et al. (2020) use this methodology on Compustat data and document an increase in the level and dispersion of firm markups. A great number of papers examine the set of assumptions needed to recover output elasticities from firm panel data (Ackerberg et al. (2015), Gandhi et al. (2020), Doraszelski & Jaumandreu (2019)). Bond et al. (2021) show that identification of output elasticity does not in general follow from revenue elasticity which is what can be inferred when using sales data only. It is still true however that the dispersion in markups is identified from the ratio estimator as long as the production elasticity of the input used is constant across firms. I impose this assumption on labour and material inputs jointly and sidestep the issue of recovering the returns to these variable inputs. Note that this is relatively innocuous in this set-up because the reallocation effects come from the dispersion in markups and not from any level effect. The most important innovation in this paper is to provide a new identification for output responsiveness which given estimates of markups also allows one to recover price pass-through.

The established approach to estimating incomplete price pass-through has been to use imported goods prices together with *exogenous* movements in exchange rates as in Goldberg & Knetter (1996), Devereux & Yetman (2010) and Gopinath & Itskhoki (2010). There is also a set of papers that estimate pass-through from tax variation either in the cross-section as in Besley & Rosen (1998) or over time (Carbonnier (2007) and Danninger & Carare (2008)). More recently, Amiti et al. (2019) use a rich dataset of Belgian exporters to estimate price pass-through with strategic settings and find that the pass-through of shocks to a firm's own cost falls with firm size as measured by employment. These results are the most directly comparable to my empirical findings given they also focus on the relationship between pass-through and firm size and it is worth noting that they point in the same direction.

3.2 Identification

To identify output responsiveness, I develop a new non-parametric method that relies only on observations of firm sales and variable costs and mild restrictions on the cost function. Before laying out the identification argument, I briefly review why the approaches used so far in the literature cannot be used here.

Pass-through Estimation

Most of the empirical literature has focused on the pass-through of cost changes to prices. Cost changes are usually defined as changes to the marginal cost of a good so that the estimand is $\frac{\partial log\, p}{\partial log\, mc}$. Given the relationship between the good's price and quantity as specified by the demand curve, an application of the chain rule shows that

$$\frac{\partial log\, p}{\partial log\, mc} = \frac{\partial log\, p}{\partial log\, x} \times \frac{\partial log\, x}{\partial log\, mc},$$

where the first term is the price elasticity. Hence, knowledge of markups and price pass-through would allow one to recover a measure of output responsiveness as given by $\frac{\partial log\, x}{\partial log\, mc}$. Note that this is not the same as the output response to a shock in the variable cost level c unless the production function displays constant returns to scale. Intuitively, when returns to scale are not constant, there is also an *endogenous* response to marginal costs due to the change of output produced. Knowledge of the elasticity of the marginal cost function would be sufficient to correct for this endogenous effect.[5]

There is a well-established tradition in the trade literature of using aggregate shocks such as exchange rate movements to estimate the cost pass-through for traded goods (Goldberg & Knetter (1996), Gopinath & Itskhoki (2010), Amiti et al. (2019)). These papers rely on prices being observable, either at the product or firm level or as price indices. This is not suitable in my work because prices are not observed in most large-scale firm data while using price indices is not revealing of the underlying distribution of firm-level responses.

[5]Given that the cost shifter c enters multiplicatively so that $mc = cv'(x)$, one can show that the following equality must hold $\frac{\partial log\, x}{\partial log\, mc} = \frac{\frac{\partial log\, x}{\partial log\, c}}{1+\epsilon_{mc}\frac{\partial log\, x}{\partial log\, c}}$.

Instead, I show that identification of output responsiveness is possible from the cross-section of firms with only an assumption on the homogeneity of the cost function. Furthermore, this argument is still valid for more general production functions that feature both variable inputs and fixed inputs like capital. The argument is similar to the one used for the identification of markups, but it leverages both the cost-minimization *and* the profit-maximization first-order conditions of the firm. Because these are static conditions, one can be agnostic about the dynamic properties of the firms problem and the approach remains valid under different specifications of the evolution of firm productivity. [6] The crucial element is that firms are symmetric monopolistic competitors with a scalar unobserved heterogeneity in their costs.

3.2.1 Derivative Estimator

I first derive the identification of Δ when labour is the only input in production.[7] Let $\{S_{it}, VC_{it}\}$ be the sales and variable costs of firm i in period t. Following the notation from the monopolistic framework with additive utility presented in Chapter 2, one can write these variables as

$$S_{it} = S(\lambda_t, x^*_{it}) \quad \text{and} \quad VC_{it} = c_{it} v(x^*_{it}),$$

where x^* is the optimal firm output which is determined from the first-order condition in equation (2.30) and hence is a function of the firm-specific cost shifter (c) and the demand index (λ) $x^*_{it} = x^*(\lambda_t, c_{it})$. Taking the total derivative of sales and variable costs with respect to the unob-

[6] The dynamic aspect would cease to be irrelevant if there is dynamic element in demand as in the habits formation model proposed by Pollak (1970) although one could argue that these factors are much more relevant for young firms or the introduction of new products and would be of second order importance in the steady state.

[7] In the model, I also allow for overhead labour that could come from the production side or from other general business needs. The only implication is that the measure of labour costs in the data should only contain the variable part of total labour.

served cost type c_{it}, we have the following equations

$$\frac{\mathrm{d}S_{it}}{\mathrm{d}c_{it}} = mr_{it} \times \frac{\partial x^*_{it}}{\partial c_{it}},$$
$$\frac{\mathrm{d}VC_{it}}{\mathrm{d}c_{it}} = v(x^*_{it}) + c_{it}v'(x^*_{it}) \times \frac{\partial x^*_{it}}{\partial c_{it}} = v(x^*_{it}) + mc_{it} \times \frac{\partial x^*_{it}}{\partial c_{it}}.$$

The cost level does not matter directly to firm sales so the effect will only show up through the response in the optimal output choice $\frac{\partial x^*_{it}}{\partial c_{it}}$. For total variable costs on the other hand, the cost level matters both directly, by changing the production costs of all units and indirectly through the output response.

Using these two expressions and the firm's profit maximization condition that equates marginal revenues to marginal costs I obtain the following expression for the relation between variable costs and sales

$$\frac{\mathrm{d}VC_{it}}{\mathrm{d}S_{it}} = 1 - (\epsilon_{v,it}\Delta_{it})^{-1}. \tag{3.3}$$

Optimality requires that a firm makes zero profits on its marginal unit so as sales change one would expect variable costs to change one-for-one and hence the constant term in equation (3.3). However, for sales to increase at a point in time it must be that the firm is moving down the demand curve, thus supplying more output and receiving a lower price. The extend of the unobserved output change will also affect the change in variable costs.

Higher output responsiveness means that a given increase in sales is concurrent with a larger output adjustment, and therefore the change in variable costs must be larger as well. On top of that, if the elasticity of variable costs is not unity then the endogenous output response kicks in a supply-side effect. In particular, if costs are locally convex ($\epsilon_{v,it} > 1$) that magnifies the total cost effect of a given output change. In the opposite case of locally concave costs, the effect will be dampening. These two cases correspond to the production function displaying decreasing returns to scale or increasing returns to scale respectively.

In identification terms, there are two unknowns $\{\epsilon_{v,it}, \Delta_{it}\}$ and only one equation. Intuitively, as we do not observe price and quantity separately, we cannot disentangle the demand side-effects that come through output responsiveness from the supply-side effects that arise from the

elasticity of variable costs. I show in Appendix **?** that if firm-level output is observed, one can estimate non-parametrically both objects. Given the data constraints, I rely on the assumption that the variable cost is homogenous of degree $1/r$ where r is known.

Estimation of Markups

Markup estimation has received growing attention in recent years with the most popular method being to exploit the cost-minimization conditions of the firm's problem. Let S denote firm sales, VC^k denote the variable expenditure on input k and r^k be the output elasticity of k. Cost minimization coupled with the assumption that the firm is a price taker in the input market implies that

$$\mu_{it} = r^k_{it} \frac{S_{it}}{VC^k_{it}}.$$

Since we observe both sales and expenditure on inputs, estimating markups boils down to estimating the elasticity of the production function. As discussed in Ackerberg et al. (2015) and Gandhi et al. (2020), this is an exercise fraught with identification problems unless more restrictions are imposed on the structure of the firm's problem. However, the production functions commonly used in the macro literature like Cobb-Douglas or CES production have the property that the elasticity r^k_{it} does not vary with the quantity that a firm produces. This is particularly relevant in this sufficient statistic approach since the reallocation channel arises from the dispersion in markups rather than the average level, which under constant cost elasticity is fully captured by the observables $\{S_{it}, VC^k_{it}\}$.

3.2.2 Estimation Framework

This subsection lists the assumptions needed for the identification of firm markup and output responsiveness as discussed above. Assumptions 1 to 3 build upon the monopolistic model presented in Chapter 2 by extending the production function of firms to allow for materials and capital. Assumption 4 introduces an ex-post shock in the price that the firm receives which is unpredictable by the firm and therefore does not

show up in the optimal output choice.[8]

Assumption 1 The production function is common up to a Hicks-Neutral productivity term ω_{it} which is known to the firm in period t

$$x_{it} = \omega_{it} F(M_{it}, L_{it}, K_{it}).$$

Assumption 2 Capital is the only fixed input that is chosen at or before $t-1$ while labour and materials are flexibly chosen at period t. Firms are price-takers in the input markets.

Assumption 3 Conditional on capital, the production function is homogenous of degree r in labour and materials, where r is known to the researcher

$$F(\theta M, \theta L, K) = \theta^r F(M, L, K) \qquad \forall M, L, K > 0 \text{ and } \lambda \geq 1.$$

The most standard case that would satisfy this restriction is Cobb-Douglas in all three inputs but other interesting functions can also be written down. An example would be a Leontief production function where capital enters separately from the variable inputs. $F(M, L, K) = \text{Min}\left(z(K), \tilde{F}(M, L)\right)$ where $z(\cdot)$ is a weakly increasing function in capital and $\tilde{F}$ is homogeneous of degree r. I show in Appendix B.1 that under Assumption 2, one can write total variable costs as

$$VC(p^M, p^L, \omega, K, x) = \mathcal{H}(p^M, p^L, K) \times \omega^{-1/r} \times x^{1/r},$$

where $\mathcal{H}$ is some function that can be solved explicitly for a given production function F. However, all one needs to establish is that input prices and the capital stock enter separately from output in the optimized cost function.

Assumption 4 Firms are profit-maximizers and face a downward-slopping inverse demand curve that is given by

$$P_{it}(x_{it}) = \lambda_t e^{\epsilon_{it}} P(x_{it}),$$

[8]This is similar to the assumption made in the production function identification literature that the unobserved Hicks neutral productivity term has a transitory component on top of a persistent one. In this instance, the iid transitory component is added to the price that the firm receives rather than its cost shifter.

where λ_t is in the period t information set of each firm while ϵ_{it} is an ex-post iid shock in the price that the firm receives and which is uncorrelated to any of the other endogenous variables. In particular, if we normalize $\mathbb{E}[e^{\epsilon_{it}}|\mathcal{I}_{it}] = \mathbb{E}[e^{\epsilon_{it}}] = 1$ we can interpret λ_t as the demand index. The slope estimator given in equation (3.3) can be extended to allow for observed heterogeneity across firms in the form of capital stock differences. Using Assumption 3 together with the fact that the profit-maximizing output choice is a static condition, one can show that the following expression holds

$$\left.\frac{\mathrm{d}VC_{it}}{\mathrm{d}S_{it}}\right|_{K_{it}} = 1 - r\left(\Delta_{it}\right)^{-1}. \tag{3.4}$$

To recover output responsiveness, one needs to estimate the same partial derivative of variable costs with respect to sales but now conditional on the capital stock of the firm. Given the iid shock in firm prices specified by Assumption 4, I use sales as the dependent variable in the estimation, with variable costs and capital as the dependent ones. Specifically, this gives the following equation for each industry-year pair

$$log(S_{it}) = m(log(VC_{it}), log(K_{it})) + \epsilon_{it}, \tag{3.5}$$

where m is some unknown function that is allowed to vary by industry and year. I run a kernel estimator on equation (3.5) and recover the elasticity of sales to variable costs and the fitted error $\hat{\epsilon}_{it}$. The mapping to the output responsiveness parameter is given by

$$\hat{\Delta}_{it} = r\left[1 - \frac{VC_{it}}{\widehat{S}_{it}}\left(\frac{\widehat{\partial\, log\, S_{it}}}{\partial\, log\, VC_{it}}\right)^{-1}\right]^{-1}. \tag{3.6}$$

Homogeneity of the production function in labour and materials also implies that all the variation in firm-level markups follows from the observed variation in sales and variable costs. I apply the ratio estimator proposed by De Loecker & Warzynski (2012) where I correct observed firm-sales for the iid shock in prices as recovered from equation (3.5) so that

$$\hat{\mu}_{it} = r\frac{\hat{S}_{it}}{VC_{it}}. \tag{3.7}$$

3.3 Data

The data used for the empirical analysis comes from the Annual Business Survey (ABS) conducted by the UK Office for National Statistics. With around 62,000 questionnaires send out every year, it is the largest firm survey in the UK and offers a very good coverage of the private sector.[9] The survey is a census of very large companies and a stratified sample of smaller ones.

Importantly, the ABS contains data on total turnover, purchases of materials and services and employment costs. Because the type and length of the questionnaire varies both by industry and firm size, the breakdown of these aggregates to more specific items is sometimes possible, however it cannot be used at large. It also contains information on capital expenditure for three different items (land & buildings, vehicles and plant & machinery) but no estimate of the capital stock of the firm. I construct the firm-level capital stock using observed investments over the years for which that particular firm is surveyed and an initial allocation rule for the first year that the firm is ever sampled. A detailed description of this procedure is available in the Online Appendix.

When applying my identification strategy to the data I have to choose what is an industry. In other words, one has to classify firms as either producing different varieties of the same good or producing distinct goods altogether which have different demand schedules as determined by the unknown industry-specific utility $u^j(\cdot)$.

To do this assignment, I use the industrial classification of each firm as recorded by the ABS. When choosing the level of industrial aggregation, one must strike a balance between ensuring that firms assigned to the same industry are not producing too different products while also having sample sizes with enough statistical power.[10] I balance these two con-

[9] The sectors that are only partially covered are either mainly publicly supplied (Education and Health) or are sectors that I exclude from my analysis given their particular features (Agriculture and Financial & Insurance activities).

[10] One would not necessarily want to use the narrowest industrial definition available even if sample size is not a concern. That is because most firms produce more than one good while they are assigned to a single subclass in the dataset. This problem is of course more serious for the very largest firms.

siderations by choosing the 2-digit level of industrial aggregation which consists of 88 different groups of industries for the UK's Standard Industrial Classification (SIC) 2007. More details on the UK's SIC07 design and firm classification can be found in the Online Appendix.

3.4 Empirical Findings

This section presents results for Manufacturing industries, while I collect results for the other five sectors in the Online Appendix. The industry-level trends I highlight below apply across sectors with very few exceptions. The objective is to show how markups, output responsiveness and price pass-through vary in the cross-section of firms and by industry.

Markups Decrease with Firm Size

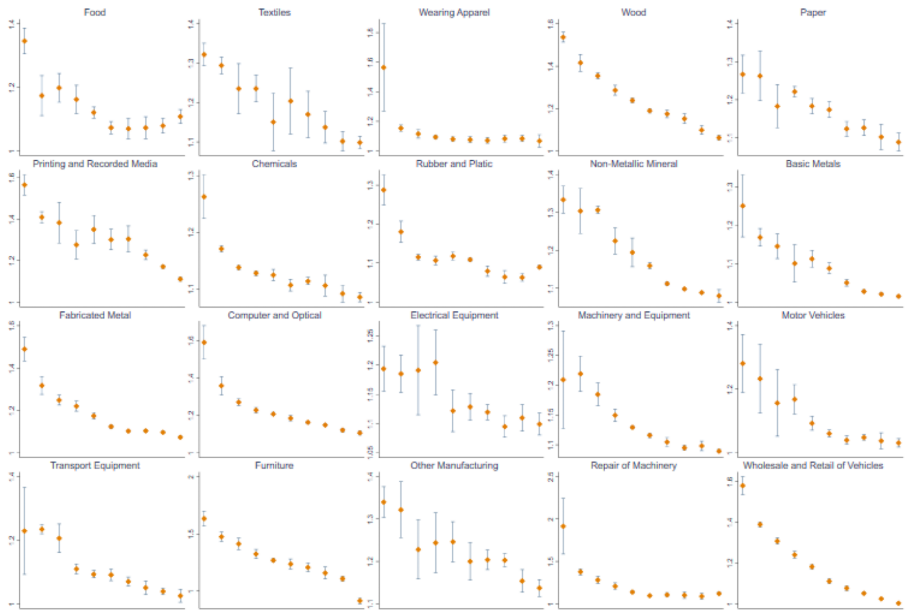


Figure 3.1: Results are ordered from the lowest to the highest sales decile as we move from left to right. from Diamonds indicate coefficient estimates of the median markup and lines indicate 95% confidence intervals obtained by bootstrapping using data for 2010. Pulling observations across years is in general not possible since firm sales also include a time fixed effect that comes from the unobserved industry-specific demand index. I use median for its robustness to outliers but using the mean gives very similar results.

Output Response Increases with Firm Size

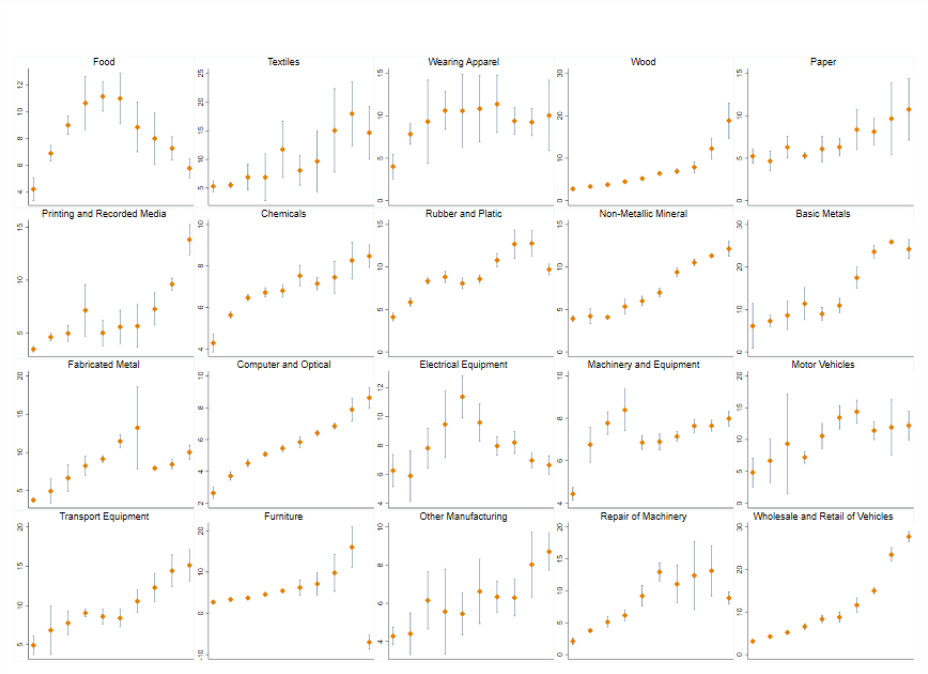


Figure 3.2: Results are ordered from the lowest to the highest sales decile as we move from left to right. Diamonds indicate coefficient estimates of the median output response and lines indicate 95% confidence intervals obtained by bootstrapping using data for 2010. Pulling observations across years is in general not possible since firm sales also include a time fixed effect that comes from the unobserved industry-specific demand index. I use median for its robustness to outliers but using the mean gives very similar results.

Finally, I plot the results for price pass-though since this is a statistic more commonly reported in other studies. From the recovered output response and markup we can back it out using the following identity

$$\frac{\partial log\, p}{\partial log\, c} = \frac{\partial log\, p}{\partial log\, x} \times \frac{\partial log\, x}{\partial log\, c} = \frac{1}{\epsilon} \times \Delta.$$

Price Pass-through Decreases with Firm Size

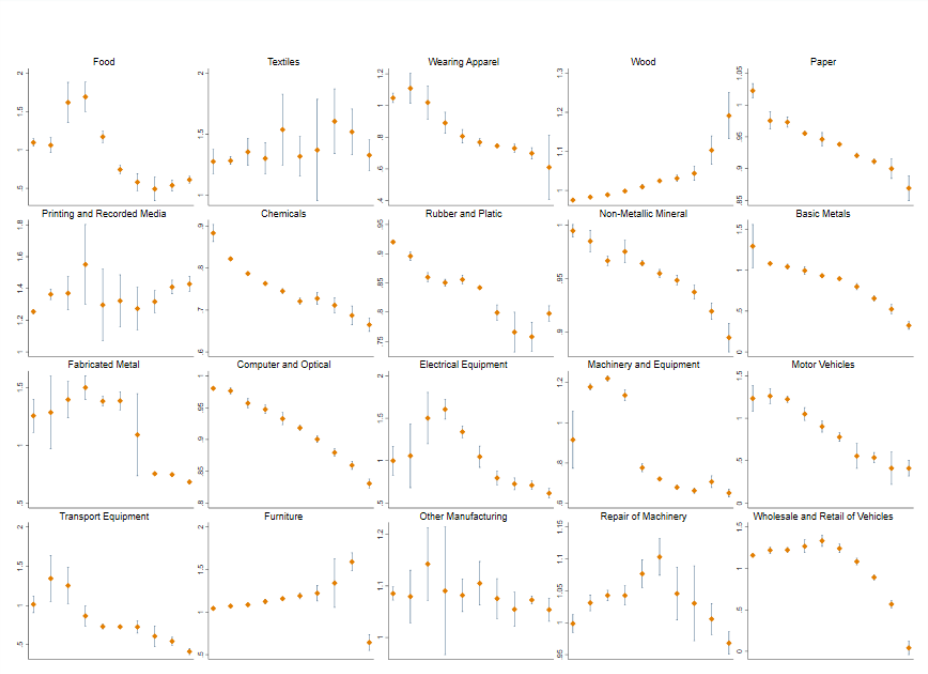


Figure 3.3: Results are ordered from the lowest to the highest sales decile as we move from left to right. Diamonds indicate coefficient estimates of the median price pass-through and lines indicate 95% confidence intervals obtained by bootstrapping using data for 2010. Pulling observations across years is in general not possible since firm sales also include a time fixed effect that comes from the unobserved industry-specific demand index. I use median for its robustness to outliers but using the mean gives very similar results.

3.4.1 Which Input Bundle?

The identification strategy in this paper relies on an assumption of conditional homogeneity for some variable input(s) in the production function. With constant elasticity of substitution between labour and materials as often assumed to be the case, either of those inputs could be used to recover markup and output responsiveness.

However, as has been pointed out by Raval (2020), using materials or labour as the variable cost measure leads to different conclusions about the size-markup relationship. [11] Note that the assumptions I make do

[11]Data availability on more than one variable input expenditure coupled with a functional form assumption of the production function implies that markup is over-identified. This over-identification has been exploited by Mertens (2020) to measure

not feature this inconsistency because they only impose that the total returns to materials and labour conditional on capital are fixed while still allowing for the elasticity of substitution to vary with capital.

No estimation can ever be assumption free so the following section presents some evidence in support of the particular supply-side assumptions made in this chapter. To fix ideas and make the exercise informative relative to the literature, I will take the Kimball aggregator that is used in Edmond et al. (2018). The Kimball aggregator represents a family of implicitly additive utilities where the utility derived from consuming a bundle $[x(i)]_{i=0}^{M}$ is implicitly given by

$$\int_0^M \Upsilon\left(\frac{x(i)}{U}\right) di = 1, \tag{3.8}$$

where Υ is a strictly increasing and strictly concave function that can be parametrized in different ways. Specifically, I will use the Klenow & Willis (2016) specification that takes that depends on two parameters $\{\sigma, \kappa\}$. The parameter σ controls the average elasticity and therefore the average markup, while the superelasticity parameter κ governs how price elasticity changes with the price level and therefore it determines the size-markup relationship. If $\kappa < 0$ we get a negative relationship of markup with firm size with demand becoming more elastic as firms sell more output, $\kappa > 0$ implies a positive size-markup relationship while $\kappa = 0$ the demand structure collapses to the CES case. Using the expression for sales and markups under the Klenow-Willis assumption, we arrive at the following relationship

$$\ln S_{it} = D_t - \frac{1}{\kappa}\left(\frac{1}{\epsilon_{it}} + ln\,\epsilon_{it}\right). \tag{3.9}$$

I estimate the above equation at the SIC2 industry level using a non linear least squares to recover the superelasticity parameter. I do this twice, once using only labour costs as variable inputs and the second time using the sum of labour and materials. Figure 3.4 plots the distribution of the parameter under both assumptions.

monopsony power in labour markets.

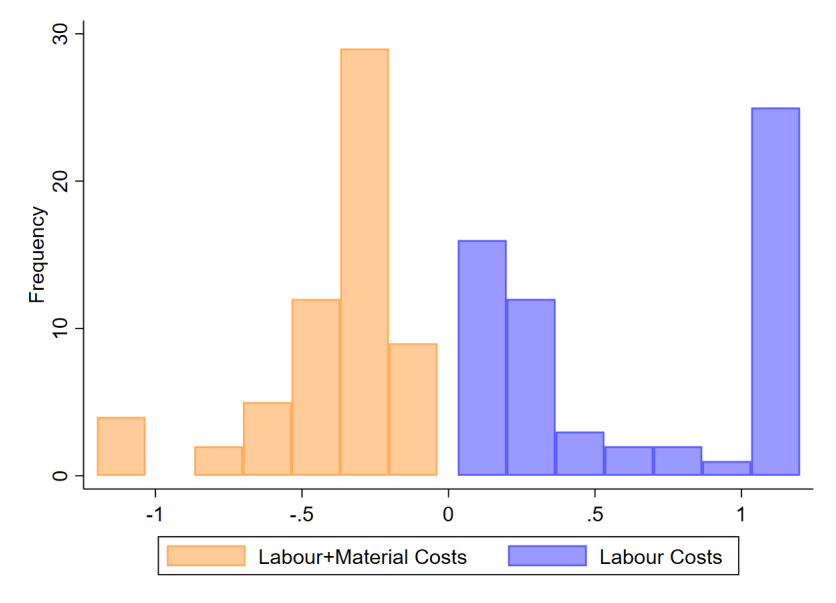


Figure 3.4: Each estimate of the superelasticity parameter corresponds to a SIC2 industry in 2010. Values that are less than -1 or larger than one have been bunched together and are shown in the two tail columns.

For all industries, the superelasticity is positive when using labour cots only which produces a positive firm size - labour marginal revenue product and is in line with what Edmond et al. (2018) document using US data. [12] There is also considerable dispersion across industries, hinting that a one sector model is unsuitable for quantifying misallocation and studying policy interventions.

On the other hand, when I use the sum of labour and materials as the variable input with constant elasticity in production, I recover a negative size-markup relationship. Again, while this discrepancy might not be well appreciated in the literature, it is a similar finding to Raval (2020) who uses manufacturing data from four different countries as well as US retail data. These results tell us that production functions like CES which feature constant output elasticities for both labour and materials are rejected by the empirical evidence, but beyond that, it remains unclear how to adjust the production function assumptions we commonly make.

[12] Their estimation of the superelasticity parameter matches their model's moment conditions to the data and so cannot be replicated without solving for the model first. Nonetheless, their benchmark calibration of 0.14 is close to the superelasticity that I recover for about half of the industries.

To make headway on this issue, I consider how well these two alternative assumptions on the output elasticity of different inputs fit the data. Figure 3.5 plots the R^2 measure for the two specifications and the very clear pattern that emerges is that using the bundle of inputs is far superior in terms of fitting the data than using labour only. Of course, this is not a test of the model itself as I am maintaining other assumptions like a Klenow-Willis demand specification, no market power in input markets, Hicks-neutral productivity differences across firms and so on. [13] Nonetheless, it provides very compelling evidence for the type of monopolistic models that we use in macro.

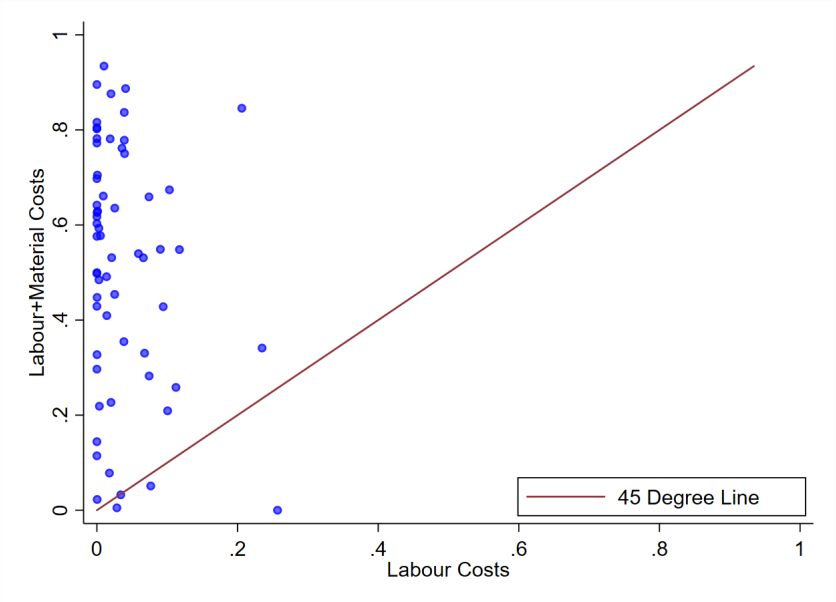


Figure 3.5: Each point corresponds to a SIC2 industry in 2010.

3.5 Tax Policy

The substantial and systematic heterogeneity in markups and output responses recovered in the cross-section of firms indicates that there is room for differential taxation to alleviate misallocation. Different tax instruments can be used to achieve this purpose by altering either the

[13]For example, Raval squares these contradictory evidence by allowing for non-neutral productivity differences across firms but imposing a constant return to scale assumption. For a more flexible estimator that does not make the latter assumption, see demirer2020production.

revenue or the cost side of the firm's problem. Examples of the first type include sales and valued added taxes while payroll taxes fall in the second category. A change in the profit tax schedule would also have first-order effects on allocative efficiency when demand is not CES and firms are heterogeneous. That is the case as long as the reform changes average post-tax profits in the economy which implies that the demand index λ must adjust. The heterogeneity in firm markups and pass-through implies that the re-allocation channel will be non-zero. However, because a profit tax does not directly affect the firm's pricing decision, it is less targeted and has less power to address allocative losses.

I will therefore pursue an application to revenue taxes. In particular, since the value added tax is relatively large and important in the UK[14], I consider the welfare implications of changing its schedule in a differential way.[15] This exercise is pertinent to many other advanced economies which also apply a valued added tax with the exception of the US, where there are sales taxes at the state level but no federal ad valorem tax. Note that in the benchmark model without materials in production, a sales and value added tax are exactly the same. With intermediary goods, they remain the same as long as the tax change pass-through for materials is complete.

The first step is to test the optimality of the observed sales tax schedule. I start by incorporating a sales tax to the original framework presented in Chapter 2 and consider the welfare incidence of reforming the tax schedule in an arbitrary way. This analysis highlights that for a constrained social planner, the *welfare weight* of each firm is a function of both the firm's markup and its output responsiveness. Intuitively, the planner has to compare the magnitude of a unit firm-specific tax shock, which in turn depends on the endogenous response of firm output. Having defined the welfare weights, I then check their empirical distribution

[14]It is the third largest tax revenue source for the UK government and accounted for about 17% of total tax receipts in $2016-2017$. The standard VAT rate has also increased over time with the latest rise happening in 2011 and consisting of an increase in the standard rate from 17.5% to 20%.

[15]Differential payroll taxation by firm size could also be an interesting application but requires more realistic assumptions on the labour market and is outside the scope of this paper.

in the data and leverage the near monotonicity with respect to sales to examine a revenue-neutral two-tier reform of the value added tax.

3.5.1 A Firm-Specific Tax

Consider the (unrealistic) case where the government has full information so that it knows the type c of each firm and can therefore charge a firm-specific linear sales tax given by $t(c)$. The first order condition of the firm is modified as following

$$(1 - t(c))\lambda(xu''(x) + u'(x)) = cv'(x). \tag{3.10}$$

This tax wedge will also show up in the cut-off and free entry condition but will not affect the resource constraint as long as it is rebated back to the household.

Let $\hat{t}$ be an arbitrary tax reform. The perturbed tax schedule is given by $t(c) + \theta\hat{t}(c)$ where $\theta \in \mathbb{R}$ parametrizes the size of the reform.

While in Chapter 2, I considered a general cost-shock that shifts the marginal cost curve of each firm, the change in the retention rate $(1-t(c))$ shifts the marginal revenue curve instead. The response in output is still governed by the elasticity of the marginal revenue and marginal cost curve at the initial equilibrium and hence

$$\frac{\hat{x}(c)}{x(c)} = [\epsilon_{mr}(x) + \epsilon_{mc}(x)]^{-1} \left(\frac{\hat{\lambda}}{\lambda} - \frac{\hat{t}(c)}{1 - t(c)} \right). \tag{3.11}$$

The incidence on the demand index is given by

$$\frac{\hat{\lambda}}{\lambda} = \int \frac{\hat{t}(c)}{1 - t(c)} \tilde{s}(c)\, dc \quad \text{where} \quad \tilde{s}(c) = \frac{(1 - t(c))s(c)g(c)}{\int (1 - t(c))s(c)g(c)\, dc}. \tag{3.12}$$

Similarly to equation 2.12, the demand index response is equal to the *average* tax shock change. The distinction is that the firm-specific shocks to the retention rate $\frac{\hat{t}(c)}{1-t(c)}$ are now weighted by the firm sales shares rather than the variable cost shares as was the case with cost shocks. Also note that to construct the correct shares one needs to use the post-tax distribution of sales. That implies that if the initial equilibrium feature a flat tax rate, pre and post-tax revenues are proportional for all firms and there is no distinction between these two measures of sales shares.

Effect on Government Revenue

The effect of a tax reform $\hat{t}$ on government revenue is determined by the equilibrium response of firms and is given by

$$\hat{\mathcal{R}}(\hat{t}) = \int \hat{t}(c)s(c)\ dG(c) + \int t(c)\hat{s}(c)\ dG(c). \tag{3.13}$$

The first term is simply the mechanical effect of changing the tax rate by $\hat{t}$. The behavioral effect of the reform goes into the second term and equals the sales response of the firm multiplied by the rate at which the government taxes the sales of that firm. Summing these effects over all firm types weighted by their density $g(c)$ gives the incidence on total tax receipts. I use equation (3.11) to rewrite the tax incidence formula in terms of firm-level elasticities so that

$$\begin{aligned}\hat{\mathcal{R}}(\hat{t}) = &\int \hat{t}(c)s(c)\ dG(c) + \int t(c)s(c)\frac{\Delta(c)}{\mu_f(c)}\frac{\hat{t}(c)}{1-t(c)}\ dG(c) \\ &+ \frac{\hat{\lambda}}{\lambda}\int t(c)s(c)\left[1+\frac{\Delta(c)}{\mu_f(c)}\right]\ dG(c).\end{aligned} \tag{3.14}$$

The behavioral effect is composed of two parts. The first one is the *partial equilibrium* effect of a firm adjusting its output and hence sales as a consequence of the tax shock it receives. The second term is due to a *general equilibrium effect* and therefore scales in the demand index response $\frac{\hat{\lambda}}{\lambda}$. One part of this effect is mechanical as tax revenues automatically increase (decrease) when aggregate demand expands (contracts) while the other part is behavioural and depends on each firm's response in exactly the same way as a tax or unit cost shock.

Effect on Aggregate Utility

The decomposition of total welfare change also features an extensive margin which is equal to $\tilde{s}_d\hat{c}_d\left[\frac{u(x_d)}{u'(x_d)x_d} - \mathcal{M}\right]$. The sign of this channel depends both on whether selection is weakened ($\hat{c}_d > 0$) or strengthened ($\hat{c}_d < 0$) following a tax change and how the consumer surplus from the smallest firm compares to the average given by $\mathcal{M}$. Because I do not recover the underlying utility function ($u(\cdot)$) it is not possible to quantify the second term and thus pin down this channel. Importantly though, the selection effect scales in the sales share of the smallest firm type

denoted by $\tilde{s}_d$. The extreme concentration of sales in the largest firms means that the difference in consumer surpluses has to be pretty large for the effects on the extensive margin to be of a similar order of magnitude to the intensive one.[16] Therefore, I will only concentrate on the direct and reallocation channel of a tax reform. Specifically, the welfare change is given by

$$\lambda \hat{U} = -\mathcal{M}\hat{\mathcal{R}} + \int \left(1 - \frac{\mathcal{M}}{\mu_f}\right) \frac{\hat{x}}{x} \tilde{s}(c)\, dc. \tag{3.15}$$

The first term is the direct effect of the perturbation in the tax schedule which is given by the total change in the tax burden $\hat{\mathcal{R}}$ weighted by the average consumer surplus. The second term is the usual reallocation channel, with firm-level output responses determined as in equation 3.11.

Elementary Tax Reform

To gain intuition, it is useful to first consider the case of changing the tax rate of a single type of firm. All possible tax reforms can be written as linear combinations of these elementary reforms. In particular, consider shocking the tax rate of type c^* by $\hat{t}(c^*) = \theta(1 - t(c^*))$ while $\hat{t}(c) = 0$ for all other firms. Notice that the particular form of the tax change expression means that θ is a shock to the retention rate.

Total Welfare Impact

Given the assumption of homogeneity of taste and income across consumers, I assume that the government evaluates social welfare just as the representative consumer does.[17] Let ψ be the marginal value of pub-

[16] The caveat is of course that the magnitude cannot be theoretically bounded so it might be erroneous to think that it is small. However, the argument on why it is reasonable to ignore the selection channel in the quantification of this model relies on an empirical fact. In other words, if the world was such that there was a high number of the smallest firms that *accounted for a non-negligible share* of total sales, one could not reasonably disregard it.

[17] In practise, the government might have other reasons for wanting to subsidise or tax deferentially by product in particular with regards to externalities that are not captured by the market price or as a means of income-redistribution. In the UK for example, education provision is exempt from VAT. Furthermore, certain goods that are considered *necessities* are taxed at a lower rates, for example food is zero-rated while domestic heating fuel is taxed at 5%. These types of considerations have been

lic funds in this economy. We can derive the total welfare impact of the elementary reform at c^* as the sum between the effects on agent's utility and on government revenue $\widehat{\mathcal{W}} = \lambda \hat{U} + \psi \hat{\mathcal{R}}$. For the elementary reform defined in the previous section, this expression evaluates to

$$\widehat{\mathcal{W}} = \tilde{s}(c^*) \left\{ - \left(\omega(c^*) - \bar{\omega} \right) + (\psi - \mathcal{M}) \hat{\mathcal{R}}(c^*) \right\}, \tag{3.16}$$

where

$$\omega(c) = \left(1 - \frac{\mathcal{M}}{\mu_f(c)} \right) \Delta(c) \quad \text{and} \quad \bar{\omega} = \int \omega(c) \tilde{s}(c) \, dc. \tag{3.17}$$

The first term is the welfare effect due to the reallocation of production away from firms of type c^* to other firms in the economy. The second term is due to differences between the marginal value of resources used by the public sector ψ and the marginal value of resources employed by the private sector λU and weighted by the amount of funds passed from private to public hands as a result of the reform. As a benchmark case, I assume that any extra funds that the government raises, will be redistributed back to private firms in a lump-sum fashion which implies that $\psi = \lambda U$ and hence the second term disappears.[18] Equation 3.16 says that there are gains in moving away from the current flat level of sales taxes as long as the distribution of $\omega(c)$ is not constant across firms.

In the special case of CES demand, the weights $\omega(c)$ are zero for all firms. Unsurprisingly, there cannot be welfare improving tax reforms if the economy is already on the Pareto frontier. For any other demand system and heterogenous firms, welfare weights are generally different from zero and potential welfare gains have to be estimated empirically or calibrated in a model. Finally, note that gains from differential taxation are not determined only by the demand side of the economy through through the shape of the utility function but also depend crucially on the

studied previously in papers such as Kopczuk (2003) and Saez (2002). They are tangential to the issue of using taxes to improve the allocative efficiency of markets and hence are better understood separately.

[18] One can think of different political economy reasons that would make this equality not true. For example, in a model with public good provision and an upper bound on tax levels, it could be the case that $\psi > \lambda U$. The question addressed here is not whether the private sector is taxed too little or too much but whether one can improve welfare for any fixed level of government revenue.

supply-side given that the distribution of firm productivities determines the demand index and the average consumer surplus in equilibrium.

Empirical Properties of ω

To evaluate a size-dependent tax change we need to know how welfare weights $\omega(c)$ vary with firm size. We know that a higher firm markup implies a higher weight as the utility derived from the marginal unit of that firm is larger. To get the total utility impact of a firm we also need to multiply by the output response parameter Δ. In the data, firm markups typically fall with size while the output responsiveness increases. These two forces push in opposite directions and therefore the slope of ω will in general depend on the calibration of the average consumer surplus. Figure 3.6 plots the mean welfare weight for all surveyed firms in 2010, and for three different values of $\mathcal{M}$.

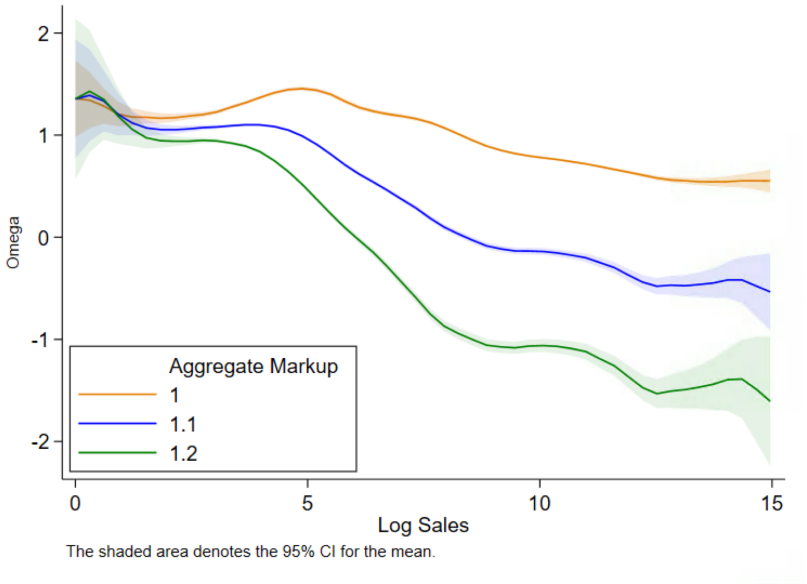


Figure 3.6: Average welfare weights by firm size for 2010.

We see that welfare weights are decreasing with firm size for all calibrations. Furthermore, this relationship becomes steeper as we increase the average consumer surplus. Intuitively, a higher $\mathcal{M}$ puts more weight on the slope of the markup-size relationship relative to the responsiveness-size one. As the first one is downward-slopping while the second one slopes upwards, this leads to a steeper decline of welfare weights with

size. Most importantly though, the slope is still negative even at the lower bound of $\mathcal{M}$. The distribution appears to be relatively tight for most firm sizes with the exception of the very smallest and the very largest of firms. Part of this comes simply due to number of observations being lower at the tails, but the dispersion among large firms also highlights that there could be motive for even more targeted tax rates.

In Figure 3.7 I plot the welfare weight-firm size relationship for each of the six sectors of the economy. Remarkably, the monotonicity of this relationships holds up pretty well for each sector as well. One pattern that emerges is that for the smallest of firms there tends to be a dip before the welfare weight picks up but because these firms collectively account for a very small proportion of sales, variations in this part of the distribution matter much less. On the other end of the firm distribution, there also appears to be an upward tick at the very end, which is most visible in the Wholesale sector, but even there the magnitude of the increase is not that large.

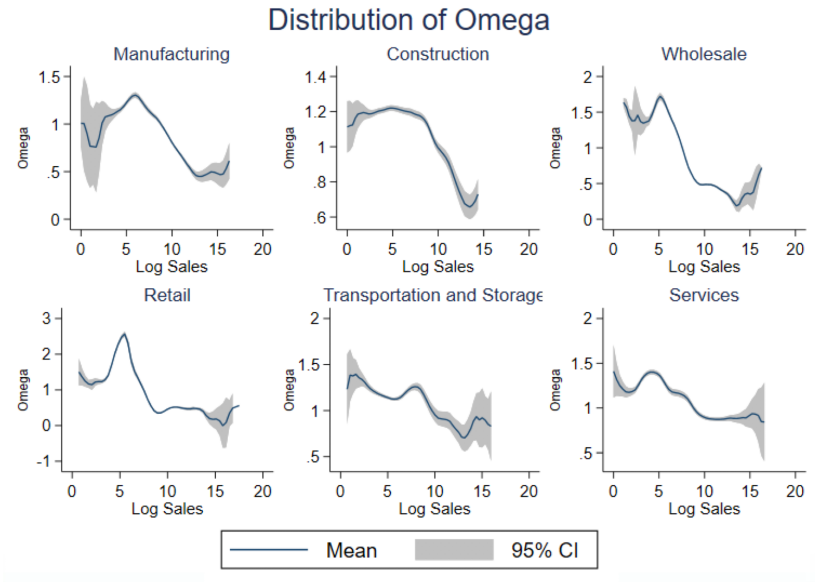


Figure 3.7: Average welfare weights by firm size for each sector in 2010.

3.5.2 A Bracket Tax Reform

One fairly simple tax reform to consider is changing the sales tax to a two step regime where the level of sales tax depends on the total sales of the firm. In other words, the idea is to pick some threshold productivity

level c^* such that after the reform the profit function of the firm is given by

$$\Pi(c) = \begin{cases} (1-\theta_1)(1-t(c))s(x) - cv(x) & c \leq c^*, \\ (1-\theta_2)(1-t(c))s(x) - cv(x) & c > c^*, \end{cases}$$

where $\{\theta_1, \theta_2\}$ are the shocks to the retention rate for firms below and above the cost-level c^* respectively. I also impose that the tax change is *revenue neutral*. This implies that total resources available to the private sector do not change and therefore welfare effects are a consequence of changes in the *production patterns* alone. Let $\{S^1, S^2\}$ denote the initial sales share of each group of firms respectively, which add up to 1 by definition. Using formula 3.14 to impose that the tax-revenue incidence of the bracket tax reform is exactly zero, I derive the following expression for the ratio of the tax shocks

$$\frac{\theta_2}{\theta_1} = -\frac{S^1}{S^2} \frac{1 + t(\bar{\Delta}_s - \bar{\Delta}_s^1)}{1 + t(\bar{\Delta}_s - \bar{\Delta}_s^2)}, \tag{3.18}$$

where

$$\bar{\Delta}_s^1 = \int_0^{c^*} \Delta_s(c)\tilde{s}(c)\, dc \quad \text{and} \quad \bar{\Delta}_s^2 = \int_{c^*}^{c_d} \Delta_s(c)\tilde{s}(c)\, dc. \tag{3.19}$$

The first term in equation 3.23 is simply the ratio of sales of the two groups while the second one is an adjustment term that shows up due to the behavioral response of firms. Naturally, if the average share-weighted sales response is the same across groups or if taxes are initially zero, the adjustment term is 1 and all that matters for balancing out tax receipts is the ratio of sales. Using the elementary tax welfare incidence as given in equation 3.16 and aggregating over all the types we get the total welfare effect of the bracket reform determined by $\{\theta_1, \theta_2, c^*\}$ is

$$\lambda \hat{U} = -\theta_1 \left\{ S^1(\bar{\omega}^1 - \bar{\omega}) + S^2 \frac{\theta_2}{\theta_1}(\bar{\omega}^2 - \bar{\omega}) \right\}, \tag{3.20}$$

where

$$\bar{\omega}^1(c) = \int_0^{c^*} \omega(c)\tilde{s}(c)\, dc \quad \text{and} \quad \bar{\omega}^2(c) = \int_{c^*}^{c_d} \omega(c)\tilde{s}(c)\, dc. \tag{3.21}$$

The welfare impact scales linearly in the size of the intervention θ_1 and it holds exactly in the limit as $\theta_1 \to 0$. Therefore, the maximum tax change impact is achieved at the cutoff level c^* that maximizes the

expression in curly brackets. To gain more intuition about the *welfare multiplier* of a revenue-neutral tax reform like this one, consider the case where the economy starts from zero sales taxes. This simplifies the ratio of the tax shocks $\left(\frac{\theta_1}{\theta_2}\right)$ to the ratio of the shares of the two groups and together with equation 3.20 gives a multiplier of

$$\hat{\mathcal{W}} = -S^1(\bar{\omega}^1 - \bar{\omega}^2).$$

There are two terms that determine what is the cut-off c^* that maximizes the welfare multiplier. Because we are shifting production from one group of firms to another by taxing the first one and subsidising the second with the proceeds, we want to maximize the difference between the $\omega(c)$ means of the two groups. Additionally, the sales share of the reference group (which in this case is the high productivity firms) given by S_1 also shows up because it determines the size of the transfer and hence how large the overall impact on the economy is.

3.5.3 Application to the UK

I investigate the welfare gains from a simple bracket tax reform like the one described above for the UK economy. To do so, I start from the observed market equilibrium for 2010 which is the last year in my sample. I assume that the tax change is economy wide and is not conditioned by sector or other observable characteristics of the firm. In particular, this implies that the extent to which the tax change hits different sectors will not be homogenous as the distribution of sales varies significantly by industry. As shown in Chapter 2, weak separability of preferences across sectors leads to a straightforward generalisation from the one sector economy. Leveraging the fact that the industry responses can be solved for independently, let a j subscript denote sector-specific variables with the demand index response given by

$$\frac{\hat{\lambda}^j}{\lambda^j} = \theta^1 S^{1j} + \theta^2 S^{2j}. \tag{3.22}$$

This formulae for the *revenue neutral* tax shock ratio in the multi-sector economy is given by

$$\frac{\theta_2}{\theta_1} = -\frac{\sum_j \alpha^j S^{1j}(1 + t[\bar{\Delta}_s^j - \bar{\Delta}_s^{1j}])}{\sum_j \alpha^j S^{2j}(1 + t[\bar{\Delta}_s^j - \bar{\Delta}_s^{2j}])}, \tag{3.23}$$

where α^j is sector's j consumption share. For any sales cut-off s^*, I can calculate for each group in each industry the sales share and average sales elasticity to solve for the ratio of shocks. Using the multi-sector results from Chapter 2, the total welfare effect is simply the sum of the industry-specific welfare changes given in equation 3.20 and weighted by the industry sales shares α^j. Finally, to make results comparable across different calibrations of the average surplus I translate the welfare measure from the money metric one to a *percentage utility change*.[19] The results from this policy experiment are shown in Figure 3.8.

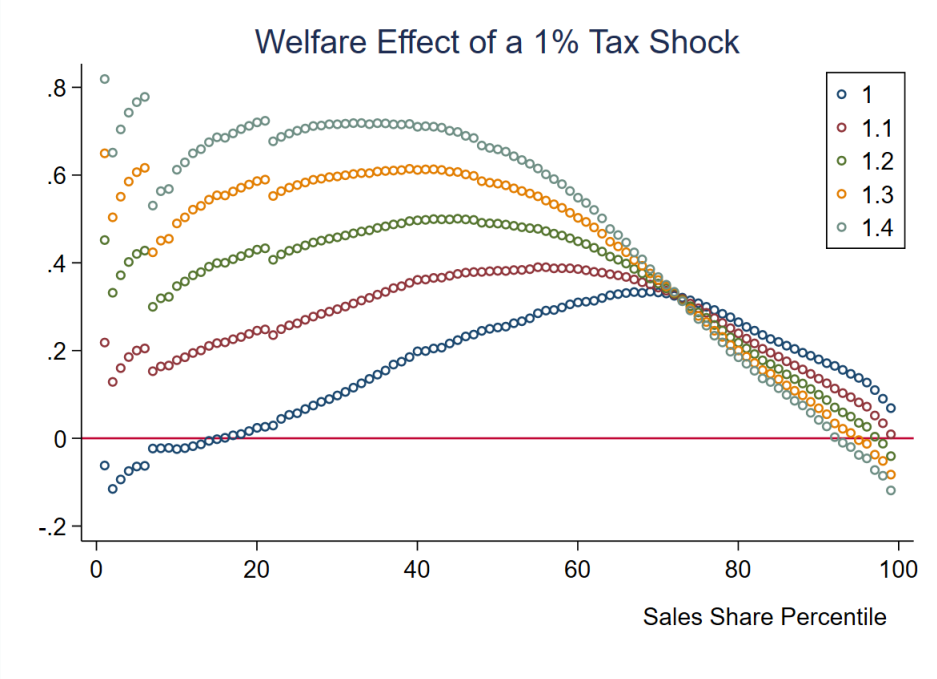


Figure 3.8: Welfare effect of a 1% tax increase on firms larger than the given sales percentile.

Discussion

The first thing to note is that the effects of increasing the VAT tax rate for *large firms* and using the proceeds to subside small firms are positive for almost all definitions of a large firm (threshold) and whatever the calibration of the average consumer surplus $\mathcal{M}$ is. A higher average sur-

[19]The money metric measure is not directly comparable for economies with different $\mathcal{M}$ because the average consumer surplus is determined by the price level in the economy. In order to convert the money metric measure in equation 3.20 to utils simply divide by $\mathcal{M}$.

plus leads to a larger maximal welfare multiplier. That is not surprising given that the slope of the welfare weights to firm sales becomes steeper as we increase $\mathcal{M}$ and so the gains from redistribution would be larger. A higher $\mathcal{M}$ also implies that the optimal sales threshold for the two-tier tax is lower.

To translate these results into a specific policy change consider increasing the VAT rate for large firms from 20% to 24% which corresponds to setting $\theta_1 = 0.05$.[20] Assuming $\mathcal{M} = 1$ as the benchmark case and choosing the $60th$ percentile of sales as our threshold we get a total welfare effect of 2%.

$$\frac{\hat{U}}{U} = 0.05 \times \textit{Welfare Multiplier} = 0.05 \times 0.4 = 0.02$$

The sales threshold for this tax reform would correspond to sales of about £2m in 2010 prices. Although this tax reform is far from eliminating all markup distortions in the market equilibrium it achieves a pretty large welfare gain.

3.6 Conclusion

This paper quantifies the gains of changing sales taxes in a non-linear way using a sufficient statistic approach that has become very prominent in the public economics literature. I develop a non-parametric estimator of firm output responsiveness by exploiting a scalar unobserved heterogeneity assumption and the monopolistic competition structure. Together with firm markup, these two objects provide the firm-level sufficient statistics needed to evaluate the gains from reallocation. This methodology is flexible and can allow for any number of industries, as well as unrestricted patterns of substitution within each industry.

I apply this method to a large dataset of UK firms and find that for almost all industries markups decrease with firm size while output responsiveness increases. With these empirical findings and my welfare formula, I evaluate the welfare gains from a VAT reform aimed at reducing misallocation. I show that a simple two-tier tax change that increases the VAT rate from 20% to 24% for firms with sales larger than £2m and

[20] Simply solve for θ such that $1 - t - \theta(1 - t) = 0.76$.

uses the proceeds to fund a VAT cut for smaller firms improves aggregate utility by about 2%. The welfare gains are robust to different calibrations of the unobserved average consumer surplus and largely support tax relief for small and medium firms at the expense of big ones.

Appendix B

Appendix to Chapter 3

B.1 Supply Side

The firm's cost minimisation problem for a given capital stock K, productivity ω and taking input prices as given is

$$\underset{M,L}{min} \quad p^M M + p^L L \quad \text{st} \quad \omega F(M, L, K) \geq Y$$

To show that the ration $\frac{L}{M}$ is independent of the total output Y, combine the FOCs with the homogeneity assumption defined in equation (3.2.2) to get

$$\begin{aligned}
\frac{p^M}{p^L} &= \frac{F_M(M, L, K)}{F_L(M, L, K)}, \\
&= \frac{F_M(M \times 1, M \times \frac{L}{M}, K)}{F_L(M \times 1, M \times \frac{L}{M}, K)}, \\
&= \frac{M^{r-1} F_M(1, \frac{L}{M}, K)}{M^{r-1} F_L(1, \frac{L}{M}, K)}, \\
&= \frac{F_M(1, \frac{L}{M}, K)}{F_L(1, \frac{L}{M}, K)}.
\end{aligned}$$

Hence we conclude that the ratio of the two variable inputs only depends on the ration of input prices and the capital stock $\frac{L^*}{M^*} = \mathcal{R}(p^M, p^L, K)$. Solving for the cost function

$$C(p^M, p^L, K, Y) = p^M M^* + p^L L^* = M^* \left(p^M + p^L \frac{L^*}{M^*} \right) = M^* \times \tilde{\mathcal{H}}(p^M, p^L, K).$$

Finally, use the fact that with positive input prices the output constraint must always be binding to solve for $M^* = \left(\frac{Y}{\omega F(1, \gamma(p^M, p^L, K), K)} \right)^{1/r}$.

Hence we conclude that for homogenous production functions and price-taking firms, the cost function is separable in output and a sub-function that depends on the input prices and the firms capital stock as following

$$C(p^M, p^L, K, Y) = \mathcal{H}(p^M, p^L, K) \times \omega^{1/r} \times Y^{1/r}.$$

B.2 Equilibrium with Taxes

Let $\mathcal{R}$ be the total revenue that the government raises from the initial sales tax $t(c)$. I assume that this tax is rebated to the household in a lump-sum fashion so that the total expenditure of the household is now $1 + \mathcal{R}$. This implies that the definition of the demand index (λ) in the equilibrium with taxes is slightly changed and is given by

$$\lambda = \frac{1 + \mathcal{R}}{M_e \int_0^{c_d} u'(x(c))x(c)\, dG(c).}$$

The equilibrium conditions are

$$\begin{aligned}
\textit{Profit Maximisation:} \quad & \lambda(1 - t(c))[u''(x(c))x(c) + u'(x(c))] = cv'(x), \\
\textit{Cut-off Condition:} \quad & \lambda(1 - t(c))[u'(x(c_d))x(c_d)] = c_d v'(x(c_d)) + f, \\
\textit{Free Entry:} \quad & \int_0^{c_d} \lambda(1 - t(c))u'(x(c))x(c) - cv(x(c)) - f\, dG(c) = f_e, \\
\textit{Government Budget:} \quad & M_e \int_0^{c_d} \lambda t(c)u'(x(c))x(c)\, dG(c) = \mathcal{R}, \\
\textit{Resource Constraint:} \quad & M_e \left(\int_0^{c_d} [cv(x(c)) + f]\, dG(c) + f_e \right) = 1.
\end{aligned}$$

B.2.1 Incidence on Tax Revenue

The total revenue that the government raises is given by

$$\begin{aligned}
\mathcal{R}+\mu\hat{\mathcal{R}} &= \int (t+\mu\hat{t})(\lambda x u' + \lambda\mu[xu''+u']\hat{x} + \mu\hat{\lambda}xu') \\
&= \int t(c)s(c) + \mu\int t(c)\lambda x u'\left(\frac{xu''+u'}{u'}\frac{\hat{x}}{x}+\frac{\hat{\lambda}}{\lambda}\right) + \mu\int \hat{t}(c)s(c) \\
&= \int t(c)s(c) + \mu\int t(c)\lambda x u'\left(\frac{1}{\mu_f}\frac{\hat{x}}{x}+\frac{\hat{\lambda}}{\lambda}\right) + \mu\int \hat{t}(c)s(c) \\
&= \int t(c)s(c) + \mu\left(\frac{\hat{\lambda}}{\lambda}\right)\int t(c)s(c)\left[1+\frac{\Delta(c)}{\mu_f}\right] \\
&- \mu\int t(c)s(c)\frac{\Delta(c)}{\mu_f}\frac{\hat{t}}{1-t} + \mu\int \hat{t}(c)s(c)
\end{aligned}$$

B.2.2 Bracket Tax Reform

Given a two-tier bracket tax reform as specified in equation (3.5.2) and an initial equilibrium with a flat sales tax denoted by t we have that

$$\begin{aligned}
\hat{R} &= \left(\frac{\hat{\lambda}}{\lambda}\right) t\int s(c)\left[1+\Delta_s(c)\right] - t\int s(c)\Delta_s(c)\theta(c) + (1-t)\int \theta(c)s(c) \\
&= \left(\frac{\hat{\lambda}}{\lambda}\right) t(1+\bar{\Delta}_s) - t[S^1\theta_1\bar{\Delta}_s^1 + S^2\theta_2\bar{\Delta}_s^2] + (1-t)[S^1\theta_1+S^2\theta_2].
\end{aligned}$$

where $\{S^1, S^2\}$ are the total sales share of the two group of firms respectively and therefore must add up to 1. The effect on the demand index is given by $\frac{\hat{\lambda}}{\lambda} = S^1\theta_1 + S^2\theta_2$ which we can substitute back into the revenue constraint expression to get

$$\begin{aligned}
[S^1\theta_1+S^2\theta_2][t(1+\bar{\Delta}_s)+1-t] &= t[S^1\theta_1\bar{\Delta}_s^1 + S^2\theta_2\bar{\Delta}_s^2], \\
[S^1\theta_1+S^2\theta_2][1+t\bar{\Delta}_s)] &= t[S^1\theta_1\bar{\Delta}_s^1 + S^2\theta_2\bar{\Delta}_s^2], \\
S^1\theta_1(1+t\bar{\Delta}_s - t\bar{\Delta}_s^1) + S^2\theta_2(1+t\bar{\Delta}_s - t\bar{\Delta}_s^2) &= 0.
\end{aligned}$$

This allows us to get an expression for the tax shock of the second group of firms which is

$$\theta_2 = -\theta_1\frac{S^1}{S^2}\frac{1+t(\bar{\Delta}_s-\bar{\Delta}_s^1)}{1+t(\bar{\Delta}_s-\bar{\Delta}_s^2)}.$$

When we have more than one sector of the economy, we need to add up the revenue effects on each sector and sum to zero. Notice that now the

sectors must be weighted by their sales share.

$$\sum_j \alpha^j \left\{ S^{1j}\theta_1(1+t[\bar{\Delta}_s^j - \bar{\Delta}_s^{1j}]) + S^{2j}\theta_2(1+t[\bar{\Delta}_s^j - \bar{\Delta}_s^{2j}]) \right\} = 0,$$

$$\sum_j \alpha^j S^{1j}(1+t[\bar{\Delta}_s^j - \bar{\Delta}_s^{1j}]) = -\frac{\theta_2}{\theta_1}\left(\sum_j \alpha^j S^{2j}(1+t[\bar{\Delta}_s^j - \bar{\Delta}_s^{2j}]) \right),$$

which is exactly the expression in equation (3.23).

B.3 When quantity is observed

As was derived in the main text, the derivative of sales to variable costs is a function of both the elasticity of the cost function $\epsilon_{v,it}$ and the output responsiveness Δ_{it}. If quantity information is also available, one can exploit it to get a second equation in $\{\epsilon_{v,it}, \Delta_{it}\}$. Taking the derivative of sales with respect to quantity we have that

$$\begin{aligned}
\frac{\mathrm{d}VC_{it}}{\mathrm{d}x_{it}} &= \frac{\mathrm{d}VC_{it}}{\mathrm{d}c_{it}}\frac{\partial c_{it}}{\partial x_{it}}, \\
&= \left(v(x_{it}^*) + mc_{it} \times \frac{\partial x_{it}^*}{\partial c_{it}} \right) \frac{\partial c_{it}}{\partial x_{it}}, \\
&= \left(v(x_{it}^*) \frac{1}{\frac{\partial x_{it}^*}{\partial c_{it}}} + mc_{it} \right) \\
&= \frac{c_{it} v(x_{it}^*)}{x_{it}} \left(-\frac{1}{\Delta_{it}} + \frac{x_{it} mc_{it}}{c_{it} v(x_{it}^*)} \right), \\
\frac{\mathrm{d}\ln VC_{it}}{\mathrm{d}\ln x_{it}} &= \left(\epsilon_{v,it} - \frac{1}{\Delta_{it}} \right).
\end{aligned}$$

Note that we are relying on the demand and cost-function being well-behaved so that the optimal output is always decreasing in firm costs and therefore the $x^*(c)$ function can be inversed. We now have a system of two equations in two unknowns and it is evident that the system can be inverted to recover the elasticity of cost and output responsiveness.

Is this sufficient for the welfare statistic? Yes, because we can now recover marginal costs and hence markups from knowledge of the elasticity of costs. Specifically, the following identity holds

$$\epsilon_{v,it} = \frac{x_{it} mc_{it}}{ac_{it}},$$

where ac_{it} denotes the average costs of firm i producing output level x_{it}.

References

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, *83*(6), 2411–2451.

Aghion, P., & Howitt, P. (1990). *A model of growth through creative destruction.* National Bureau of Economic Research Cambridge, Mass., USA.

Amiti, M., Itskhoki, O., & Konings, J. (2019). International shocks, variable markups, and domestic prices. *The Review of Economic Studies*, *86*(6), 2356–2402.

Arkolakis, C., Costinot, A., Donaldson, D., & Rodríguez-Clare, A. (2019). The elusive pro-competitive effects of trade. *The Review of Economic Studies*, *86*(1), 46–80.

Atkeson, A., & Burstein, A. (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review*, *98*(5), 1998–2031.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & Van Reenen, J. (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, *135*(2), 645–709.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., Van Reenen, J., et al. (2017). *The fall of the labor share and the rise of superstar firms.* National Bureau of Economic Research.

Baqaee, D. R., & Farhi, E. (2020). Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics*, *135*(1), 105–163.

Barkai, S. (2016). Declining labor and capital shares. *Stigler Center for the Study of the Economy and the State New Working Paper Series*, *2*.

Barkai, S. (2020). Declining labor and capital shares. *The Journal of Finance*, *75*(5), 2421–2463.

Basu, S., & Fernald, J. G. (1997). Returns to scale in us production: Estimates and implications. *Journal of political economy*, *105*(2), 249–283.

Benassy, J.-P. (1996). Taste for variety and optimum production patterns in monopolistic competition. *Economics Letters*, *52*(1), 41–47.

Besley, T. J., & Rosen, H. S. (1998). *Sales taxes and prices: an empirical analysis* (Tech. Rep.). National Bureau of Economic Research.

Bilbiie, F. O., Ghironi, F., & Melitz, M. J. (2012). Endogenous entry, product variety, and business cycles. *Journal of Political Economy*, *120*(2), 304–345.

Blonigen, B. A., & Pierce, J. R. (2016). *Evidence for the effects of mergers on market power and efficiency* (Tech. Rep.). National Bureau of Economic Research.

Bond, S., Hashemi, A., Kaplan, G., & Zoch, P. (2021). Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data. *Journal of Monetary Economics*.

Bulow, J. I., & Pfleiderer, P. (1983). A note on the effect of cost changes on prices. *Journal of political Economy*, *91*(1), 182–185.

Carbonnier, C. (2007). Who pays sales taxes? evidence from french vat reforms, 1987–1999. *Journal of Public Economics*, *91*(5-6), 1219–1229.

Chernozhukov, V., & Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, *115*(2), 293–346.

Danninger, M. S., & Carare, M. A. (2008). *Inflation smoothing and the modest effect of vat in germany* (No. 8-175). International Monetary Fund.

Decker, R. A., Haltiwanger, J. C., Jarmin, R. S., & Miranda, J. (2018). *Changing business dynamism and productivity: Shocks vs. responsiveness* (Tech. Rep.). National Bureau of Economic Research.

De Loecker, J., & Eeckhout, J. (2017). *The rise of market power and the macroeconomic implications* (Tech. Rep.). National Bureau of Economic Research.

De Loecker, J., Eeckhout, J., & Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, *135*(2), 561–644.

De Loecker, J., & Warzynski, F. (2012). Markups and firm-level export status. *American economic review*, *102*(6), 2437–71.

Devereux, M. B., & Yetman, J. (2010). Price adjustment and exchange rate pass-through. *Journal of International Money and Finance*, *29*(1), 181–200.

Dhingra, S., & Morrow, J. (2019). Monopolistic competition and optimum product diversity under firm heterogeneity. *Journal of Political Economy*, *127*(1), 196–232.

Dixit, A. K., & Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American economic review*, *67*(3), 297–308.

Doraszelski, U., & Jaumandreu, J. (2019). Using cost minimization to estimate markups.

Edmond, C., Midrigan, V., & Xu, D. Y. (2018). *How costly are markups?* (Tech. Rep.). National Bureau of Economic Research.

Feenstra, R. C. (2003). A homothetic utility function for monopolistic competition models, without constant price elasticity. *Economics Letters*, *78*(1), 79–86.

Feenstra, R. C. (2018). Restoring the product variety and pro-competitive gains from trade with heterogeneous firms and bounded productivity. *Journal of International Economics*, *110*, 16–27.

Gandhi, A., Navarro, S., & Rivers, D. A. (2020). On the identification of gross output production functions. *Journal of Political Economy*, *128*(8), 2973–3016.

Goldberg, P. K., & Knetter, M. M. (1996). *Goods prices and exchange rates: what have we learned?* National Bureau of Economic Research Cambridge, Mass., USA.

Gopinath, G., & Itskhoki, O. (2010). Frequency of price adjustment and pass-through. *The Quarterly Journal of Economics*, *125*(2), 675–727.

Grossman, G. M., & Helpman, E. (1991). Quality ladders in the theory of growth. *The Review of economic studies*, *58*(1), 43–61.

Gutiérrez, G., & Philippon, T. (2017). *Declining competition and investment in the us* (Tech. Rep.). National Bureau of Economic Research.

Hall, R. E. (1988). The relation between price and marginal cost in us industry. *Journal of political Economy*, *96*(5), 921–947.

Hall, R. E. (2018). *New evidence on the markup of prices over marginal costs and the role of mega-firms in the us economy* (Tech. Rep.). National Bureau of Economic Research.

Harberger, A. C. (1954). Monopoly and resource allocation. *The American Economic Review*, *44*(2), 77–87.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica: Journal of the Econometric Society*, 1127–1150.

Hsieh, C.-T., & Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, *124*(4), 1403–1448.

Karabarbounis, L., & Neiman, B. (2013). The global decline of the labor share. *The Quarterly Journal of Economics*, *129*(1), 61–103.

Karabarbounis, L., & Neiman, B. (2018). *Accounting for factorless income* (Tech. Rep.). National Bureau of Economic Research.

Kehrig, M., & Vincent, N. (2017). Growing productivity without growing wages: The micro-level anatomy of the aggregate labor share decline.

Kehrig, M., & Vincent, N. (2021). The micro-level anatomy of the labor share decline. *The Quarterly Journal of Economics*, *136*(2), 1031–1087.

Klenow, P. J., & Willis, J. L. (2016). Real rigidities and nominal price changes. *Economica*, *83*(331), 443–472.

Klette, T. J., & Kortum, S. (2004). Innovating firms and aggregate innovation. *Journal of political economy*, *112*(5), 986–1018.

Kopczuk, W. (2003). A note on optimal taxation in the presence of externalities. *Economics Letters*, *80*(1), 81–86.

Melitz, M. J., & Ottaviano, G. I. (2008). Market size, trade, and productivity. *The review of economic studies*, *75*(1), 295–316.

Mertens, M. (2020). *Micro-mechanisms behind declining labor shares: Rising market power and changing modes of production* (Tech. Rep.). Mimeo.

Montagna, C. (2001). Efficiency gaps, love of variety and international trade. *Economica*, *68*(269), 27–44.

Mrázová, M., & Neary, J. P. (2017). Not so demanding: Demand structure and firm behavior. *American Economic Review*, *107*(12), 3835–74.

Mrázová, M., Neary, J. P., & Parenti, M. (2021). Sales and markup dispersion: theory and empirics. *Econometrica*, *89*(4), 1753–1788.

Perla, J. (2015). Product awareness, industry life cycles, and aggregate profits. *University of British Columbia mimeograph*.

Peters, M. (2012). Heterogeneous mark-ups and endogenous misallocation. *Yale University manuscript*.

Pollak, R. A. (1970). Habit formation and dynamic demand functions. *Journal of political Economy*, *78*(4, Part 1), 745–763.

Pugsley, B. W., & Şahin, A. (2015). Grown-up business cycles. *The Review of Financial Studies*.

Raval, D. (2020). Testing the production approach to markup estimation. *Available at SSRN 3324849*.

Restuccia, D., & Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics*, *11*(4), 707–720.

Rognlie, M. (2016). Deciphering the fall and rise in the net capital share: accumulation or scarcity? *Brookings papers on economic activity*, *2015*(1), 1–69.

Rossi-Hansberg, E., Sarte, P.-D., & Trachter, N. (2018). *Diverging trends in national and local concentration* (Tech. Rep.). National Bureau of Economic Research.

Saez, E. (2002). The desirability of commodity taxation under non-linear income taxation and heterogeneous tastes. *Journal of Public Economics*, *83*(2), 217–230.

Traina, J. (2018). Is aggregate market power increasing? production trends using financial statements.

Vives, X. (1999). *Oligopoly pricing: old ideas and new tools*. MIT press.

Weyl, E. G., & Fabinger, M. (2013). Pass-through as an economic tool: Principles of incidence under imperfect competition. *Journal of Political Economy*, *121*(3), 528–583.

Zhelobodko, E., Kokovin, S., Parenti, M., & Thisse, J.-F. (2012). Monopolistic competition: Beyond the constant elasticity of substitution. *Econometrica*, *80*(6), 2765–2784.