Correcting Market Power with Taxation

A Sufficient Statistic Approach

Dajana Xhani [*]

December 10, 2022

Please click here for the latest version

Abstract

This paper investigates the potential for tax policy to reduce distortions caused by market power. I provide a novel non-parametric approach that does not rest on strong assumptions about demand curves while simultaneously accounting for general equilibrium effects. To that end, I derive the welfare incidence of general shocks in models of monopolistic competition and heterogenous firms. I decompose the welfare effect into three channels: (i) the direct effect of the shock, (ii) a selection effect that arises on the extensive margin and, (iii) a reallocation effect as production shifts across firms. The latter depends on the joint distribution of firm-level markups, output responsiveness and sales. I show that it is possible to recover output responsiveness non-parametrically from revenue and cost data when the production function is homogenous in variable inputs. I apply this method to a large dataset of UK firms and find that at the industry level markups are decreasing with firm size while output responsiveness is increasing. Finally, I use these results to empirically evaluate a tax reform aimed at reducing misallocation. I estimate that a simple two-tier VAT tax change that increases the VAT rate from 20% to 24% for firms with sales larger than £2m and uses the proceeds to fund a cut for smaller firms improves aggregate utility by 2%.

[*] I am indebted to my advisors, Vincent Sterk, Victor Rios-Rull and Morten O. Ravn for their invaluable guidance and support. For helpful discussions and comments, I thank Kirill Borusyak, Alan Olivi, Aureo De Paula, Lars Nesheim and Franck Portier. I thank the UK Data Service for guidance and assistance with regards to the data. The UK Data Service agrees that the outputs are non-disclosive, and cannot be used to identify a person or organisation.

University College London, D.Xhani@tilburguniversity.edu

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 decision 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.

I demonstrate that one can dispense with these parametric restrictions while staying in the class of monopolistic competition models. In particular, I derive an analytical expression for evaluating the welfare gains of arbitrary tax changes and non-parametrically estimate the firm-level sufficient statistics that appear in that formula. With these empirical findings and my welfare formula at hand, I evaluate the welfare gains for a simple 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.

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 a level difference in production costs. Profit maximization implies that more productive firms will be larger as their lower cost drives 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. To account for non-convexities in production, I also allow for fixed operating costs and a sunk cost in creating a new firm.

I solve for the first-order perturbation of the equilibrium response following a general shock to the cost distribution. 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). The firm 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. 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. 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 use the structure provided by symmetric monopolistic competition with Hicks-neutral productivity differences to derive a novel identification of output responsiveness. In particular, assume that variable inputs display fixed returns to scale in production denoted by r. Exploiting the scalar unobservable assumption in firm-level costs one can show that

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

The intuition for Equation 1 is as following. Optimality requires that a firm makes zero profits on the marginal unit sold so as sales change we expect variable costs to change one-for-one and hence the constant term. However, for sales to increase at a point in time it must be that the firm is moving down the demand curve. A higher output responsiveness means that a given increase in sales is concurrent with a larger output adjustment and so variable costs must increase faster. On top of that, if returns to scale are not constant ($r \neq 1$), an *endogenous* supply-side effect kicks in. Specifically, for a given output response, faster decreasing returns to scale mean that variable costs must increase by more. In the empirical application, I assume knowledge of the returns to scale parameter.

The empirical results are presented and discussed in detail in Section 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 in the UK dataset using either labour or the bundle of labour and materials as the variable input.[1] Secondly, I find that output responsiveness increase with firm size.

[1] 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.

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 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 price pass-through to a cost shock is generally decreasing with firm size. These results are in line with the findings in Amiti et al. (2019) who use a rich Belgian dataset of exporting firms.

In Section 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 the firm whose sales tax is increased chooses to supply less and therefore uses less resources. The labour that is freed up will be employed by other firms in equilibrium so 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) welfare 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

$$\omega = \left(1 - \frac{\mathcal{M}}{\mu_f}\right) \Delta \tag{2}$$

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, it 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 and 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 $1997 - 2010$. I use the approximate monotonicity of ω

[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.

against sales 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 point we choose to partition the firms. In any case, I find that the multiplier is positive for a large set of sale cut-offs and for any value of $\mathcal{M}$ larger than 1. In this sense, this policy change 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 large firms.

Related Literature

This paper is motivated by the recent literature on increasing firm concentration and falling labour share. Methodologically, it is related to the empirical literature on estimating markups and price pass-through at the firm or product level as well as the more theoretical 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 (Autor et al. (2020)), TFP (Decker et al. (2018)) or markups (De Loecker et al. (2020)). The reallocation of production and rise in concentration has also been documented by (Rossi-Hansberg et al. (2018), Kehrig & Vincent (2021)) while (Gutiérrez & Philippon (2017)) provide evidence that it 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.

[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 $\theta \to 0$.

(2021) show that identification of output elasticity does not in general follow from revenue elasticity which is what we can infer when using only sales data. It remains true 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 the bundle of labour and materials and sidestep the issue of recovering the elasticity by assuming instead that its value is known. I leverage the assumption of symmetric monopolistic competition across firms to provide a new identification for output response and therefore price pass-through.

A well-established approach in estimating incomplete price pass-through has been to use imported goods prices and *exogenous* movements in exchange rates (Goldberg & Knetter (1996), Devereux & Yetman (2010), Gopinath & Itskhoki (2010)). There is also a set of papers that estimate pass-through from tax variation (Besley & Rosen (1998), Carbonnier (2007), Danninger & Carare (2008)). Amiti et al. (2019) use a rich dataset of Belgian exporters to estimate price pass-through in strategic settings and for different types of shocks.

The misallocation literature started with Harberger (1954) with more recent examples including Restuccia & Rogerson (2008) and Hsieh & Klenow (2009). The standard CES assumption of Dixit & Stiglitz (1977) has to a large extent *hidden* the issue of misallocation in macro models because as Dhingra & Morrow (2019) prove, CES is the only parametrization of utility where the social planner's solution coincides to the market outcome. 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), Weyl & Fabinger (2013) Zhelobodko et al. (2012)). While these papers provide important theoretical insights they often rely on imposing further conditions on utility or analysing cases of identical firms. Conclusions about the direction and size of misallocation with firm heterogeneity are in general not possible without further restrictions.

Overall, this paper is mostly related to Edmond et al. (2018) and Baqaee & Farhi (2020). It generalized Edmond et al. (2018) by not putting any parametric restrictions on the demand function or the distribution of firm types. Unlike Edmond et al. (2018) however, it does not incorporate a truly dynamic model with endogenous capital choice so it can only capture static gains. Comparing to their conclusions, my results suggest that a Kimball calibration with positive superelasticity does not match well the microdata. On top of that, the markup-size relationship varies by industry so it is important to allow for industry-specific demand schedules even when using a parametric family. The sufficient statistic approach in this paper is very close in spirit to 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 one to consider a general input-output structure in the economy and still get a closed-form statistic for the distance to the frontier. However, this means that we are effectively discarding the information content in firm markups and that the framework is not well-suited for policy counterfactuals.

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 static one-sector model of the economy. This will highlight the firm-level objects one needs to recover from the data. Subsection 2.5 presents the multi-sector version of the model and shows that aggregating across sectors is still tractable in this framework. In the Appendix, I also consider how the welfare incidence formula changes when one relaxes some of the baseline assumptions. In particular, I consider extensions of the model by allowing for general love-of-variety, materials used in production and an endogenous labour choice.

2.1 Initial Equilibrium

2.1.1 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:

$$\underset{[x_i]\geq 0}{max} \quad \int_0^M u(x_i)\ di \qquad subject\ to \qquad \int p_i x_i\ di \leq E \tag{3}$$

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 condition to the consumer's problem gives the inverse demand

[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

function $p_i = \lambda u'(x_i)$ where λ is equal to

$$\lambda = \left(\int_0^M u'(x_i) x_i \, di \right)^{-1} . \tag{4}$$

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.

2.1.2 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

$$\underset{x}{max} \quad \lambda u'(x)x - cv(x) - f \tag{5}$$

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 (5). These assumptions are the same as in Hopenhayn (1992). Free entry implies that expected profits are equal to the sunk entry cost f_e.

2.1.3 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

ex-post profits.

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 (6) to (9).

$$\textit{Profit Maximisation:} \quad \lambda[u''(x(c))x(c) + u'(x(c))] = cv'(x), \tag{6}$$

$$\textit{Cut-off Condition:} \quad \lambda[u'(x(c_d))x(c_d)] = c_d v'(x(c_d)) + f, \tag{7}$$

$$\textit{Free Entry:} \quad \int_0^{c_d} \lambda u'(x(c))x(c) - cv(x(c)) - f \; dG(c) = f_e, \tag{8}$$

$$\textit{Resource Constraint:} \quad M_e \left(\int_0^{c_d} [cv(x(c)) + f] \; dG(c) + f_e \right) = 1. \tag{9}$$

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.

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

$$\epsilon(x) \equiv -\frac{u'(x)}{xu''(x)}, \qquad \text{and} \qquad \rho(x) \equiv -\frac{xu'''(x)}{u''(x)}, \tag{10}$$

[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.

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

$$\epsilon_{vc}(x) \equiv \frac{xv'(x)}{v(x)}, \quad \text{and} \quad \epsilon_{mc}(x) \equiv \frac{xv''(x)}{v'(x)}. \tag{11}$$

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.

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 $\hat{U}$.

[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.

2.3.1 Output Response

The firm-level output response is the solution to the perturbed profit maximisation condition in equation (6), 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

$$\frac{\hat{x}(c)}{x(c)} = \Delta(x)\left(\frac{\hat{\lambda}}{\lambda} - \frac{\hat{c}}{c}\right) \tag{12}$$

The shock shifts the marginal cost (mc) 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 (mr) 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.

2.3.2 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 (7):

$$\frac{\hat{c}_d}{c_d} = \mu_d \epsilon_{vc}(x_d)\frac{\hat{\lambda}}{\lambda} \tag{13}$$

Because the first two terms in equations (13) 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.

2.3.3 Demand Index Response

I obtain the endogenous response of the demand index from the first-order perturbation of the free-entry equation in (8). 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

$$\frac{\hat{\lambda}}{\lambda} = \frac{\text{Aggregate Variable Costs}}{\text{Aggregate Sales}} \times \int_0^{c_d} \tilde{v}(c) \frac{\hat{c}}{c}\, dc \tag{14}$$

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 (12) we can see that the output produced by each firm would remain unchanged.

2.3.4 Mass of Entrants Response

The change in the mass of entrants is derived from the resource constraint (9) which is essentially a labour market clearing condition. Specifically, if more labour is used for

[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.

production then fewer firms will be created in equilibrium. In the Appendix, I show that the entry response is given by

$$\frac{\hat{M_e}}{M_e} = -M_e \left(\hat{c}_d (c_d v(x_d) + f) g(c_d) + \int_0^{c_d} cv(x) \left(\frac{\hat{c}}{c} + \epsilon_{vc}(x) \frac{\hat{x}(c)}{x(c)} \right) \, dG(c) \right) \tag{15}$$

The two terms in equation (15) 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 initial levels of output, and a reallocation channel as firms optimally choose to adjust their output levels.

2.4 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

$$\hat{U} = M_e \hat{u} + \hat{M_e} u \tag{16}$$

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 (16) 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 $\Delta Income \times \text{MU}_{Income} \approx \Delta U$.

2.4.1 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

$$\lambda \hat{U} = \int_0^{c_d} \tilde{s}(c) \frac{\hat{x}(c)}{x(c)} \, dc = \int_0^{c_d} \tilde{s}(c) \Delta(c) \left(\frac{\hat{\lambda}}{\lambda} - \frac{\hat{c}}{c} \right) \, dc \tag{17}$$

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 $\Delta(c)$.

2.4.2 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 (17). 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

$$\mathcal{M} \equiv \lambda U = \frac{\int_0^{c_d} u(x(c)) \, dG(c)}{\int_0^{c_d} u'(x(c))x(c) \, dG(c)} \tag{18}$$

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

[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

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.

2.4.3 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

$$\begin{aligned}\lambda\hat{U} = & \overbrace{-\mathcal{M}\, M_e \int_0^{c_d} cv(x)\frac{\hat{c}}{c}dG(c)}^{\textit{direct effect}} \quad + \quad \overbrace{M_e\hat{c}_d g(c_d) s_d\left[\frac{u(x_d)}{u'(x_d)x_d} - \mathcal{M}\right]}^{\textit{selection}} \\ & + \quad \overbrace{M_e \int_0^{c_d}\left[1 - \frac{\mathcal{M}}{\mu_f(c)}\right] s(c)\frac{\hat{x}(c)}{x(c)}dG(c)}^{\textit{reallocation}}\end{aligned} \tag{19}$$

To fix ideas, consider a general cost-push shock so that $\hat{c}$ is positive for all firms. The first term in (19) 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 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.

2.4.4 A Special Case: CES Demand

With CES demand, equation (19) 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 (18) one can show that whatever the underlying distribution of firms, aggregate

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 $M_e s_d g(c_d) = \frac{M_e s_d g(c_d)}{M_e \int_0^{c_d} s(c)\ dG(c)} = \tilde{s}_d$.

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 (19) 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.

2.5 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 than 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 the household's maximization problem is given by

$$\underset{[x_i^j]_{i\in I}}{max} \quad \mathcal{F}(U^1, U^1, \dots U^k) \qquad subject\ to \qquad \sum_{j=1}^{k} M_e^j \int p^j(c)x^j(c)\, dG^j(c) \leq 1 \tag{20}$$

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 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

$$\max_{\{\alpha^1,\alpha^2,\ldots,\alpha^k\}} \mathcal{F}\left(\alpha^1 \frac{u^1}{s^1}, \alpha^2 \frac{u^2}{s^2}, \ldots, \alpha^k \frac{u^k}{s^k}\right) \quad \text{st} \quad \sum_j \alpha^j = 1 \tag{21}$$

where the FOC requires that

$$\mathcal{F}'_j \frac{u^j}{s^j} - \frac{1}{\psi} = 0 \tag{22}$$

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

$$\psi \hat{U} = \sum_{j=1}^{k} \alpha^j \left[\frac{\hat{\alpha}^j}{\alpha^j} + \frac{\hat{u}^j}{u^j} - \frac{\hat{s}^j}{s^j}\right] \tag{23}$$

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 (19) 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.

3 Identification

Having derived the non-parametric welfare formula, I now discuss how to recover the firm-level markup and output responsiveness which are the two sufficient statistics that enter in the reallocation channel as specified by equation (19).

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

[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 the Appendix.

I estimate firm-level markups using the ratio estimator developed by De Loecker & Warzynski (2012). 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-passthrough 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.[15]

There is a well-established tradition in the trade literature of using aggregate shocks such as exchange rate movements to estimate the cost-passthrough 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.

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, I can remain agnostic about the dynamic properties of the firms problem and the

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

approach remains valid under different specifications of the evolution of firm productivity. The crucial element is that firms are symmetric monopolistic competitors with a scalar unobserved heterogeneity in their costs.

3.1 Derivative Estimator

I first derive the identification of Δ when labour is the only input in production.[16] Let $\{S_{it}, VC_{it}\}$ be the sales and variable costs of firm i in period t. Given the assumptions presented in section 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 (6) so that we can write it as $x^*_{it} = x^*(\lambda_t, c_{it})$. Taking the total derivative of sales and variable costs with respect to the unobserved cost type c_{it}, we have the following equations

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

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

$$\frac{\mathrm{d}VC_{it}}{\mathrm{d}S_{it}} = 1 - (\epsilon_{vc,it}\Delta_{it})^{-1} \tag{24}$$

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 (24). 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

[16] 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 the total labour used.

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 ($\epsilon_{vc,it} > 1$) 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 $\{\epsilon_{vc,it}, \Delta_{it}\}$ 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 the 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 that input 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, for the set of production functions often used in the macro literature like Cobb-Douglas or CES production, the elasticity r^k_{it} is a constant. This is particularly relevant in my sufficient statistic approach since the reallocation channel arises from the dispersion in markups rather than the average level, which in this instance is fully captured by the observables $\{S_{it}, VC^k_{it}\}$.

3.2 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 Section 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.[17]

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. $F(M, L, K) = \text{Min}\Big(z(K), \tilde{F}(M, L)\Big)$ where $z(\cdot)$ is a weakly increasing function in capital and $\tilde{F}$ is homogeneous of degree r. I show in the appendix that under Assumption 2, we 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*

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

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.

[17] 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.

The slope estimator given in equation (24) 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

$$\left.\frac{\mathrm{d}VC_{it}}{\mathrm{d}S_{it}}\right|_{K_{it}} = 1 - r\left(\Delta_{it}\right)^{-1} \tag{25}$$

To recover output responsiveness, we need an estimate of the partial derivative of variable costs with respect to sales, conditional on capital. 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

$$log(S_{it}) = m(log(VC_{it}), log(K_{it})) + \epsilon_{it} \tag{26}$$

where m is some unknown function that is allowed to vary by industry and year. I run a kernel estimator on Equation (26) 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

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

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 (26).

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

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-passthrough vary in the cross-section of firms and by industry. Firm markup and responsiveness feed into the welfare formula (19) and hence the distribution of these sufficient statistics governs the strength of the reallocation channel for different firm sizes.

4.1 Data description

The data used for the empirical analysis comes from the Annual Business Survey (ABS) conducted by the Office for National Statistics. It is the largest such survey in the UK with $62,000$ questionnaires send out every year and with great coverage of the private sector.[18] The survey is a census of very large companies and a stratified sample of smaller ones. Up to 1997, it only covered production and construction so I focus on the later part of the sample.

Importantly, the survey 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 Data 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. 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 and having sample sizes with enough statistical power.[19] I balance these two considerations 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 appendix.

[18] 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).

[19] 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.

Markups Decrease with Firm Size

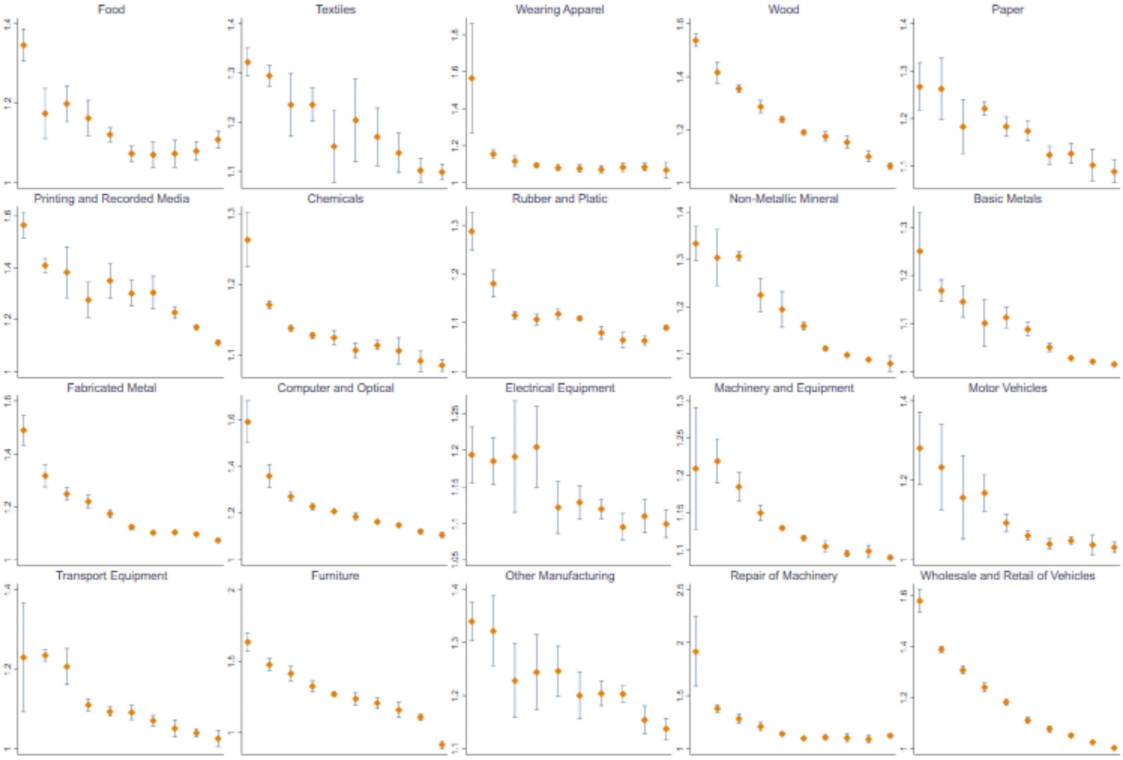


Figure 1: Results are ordered from the lowest to the highest sales decile as we move from left to right. from Diamonds indicate coefficient estimates of the median markup and lines indicate 95% confidence intervals obtained by bootstrapping using data for 2010. Pulling observations across years is in 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

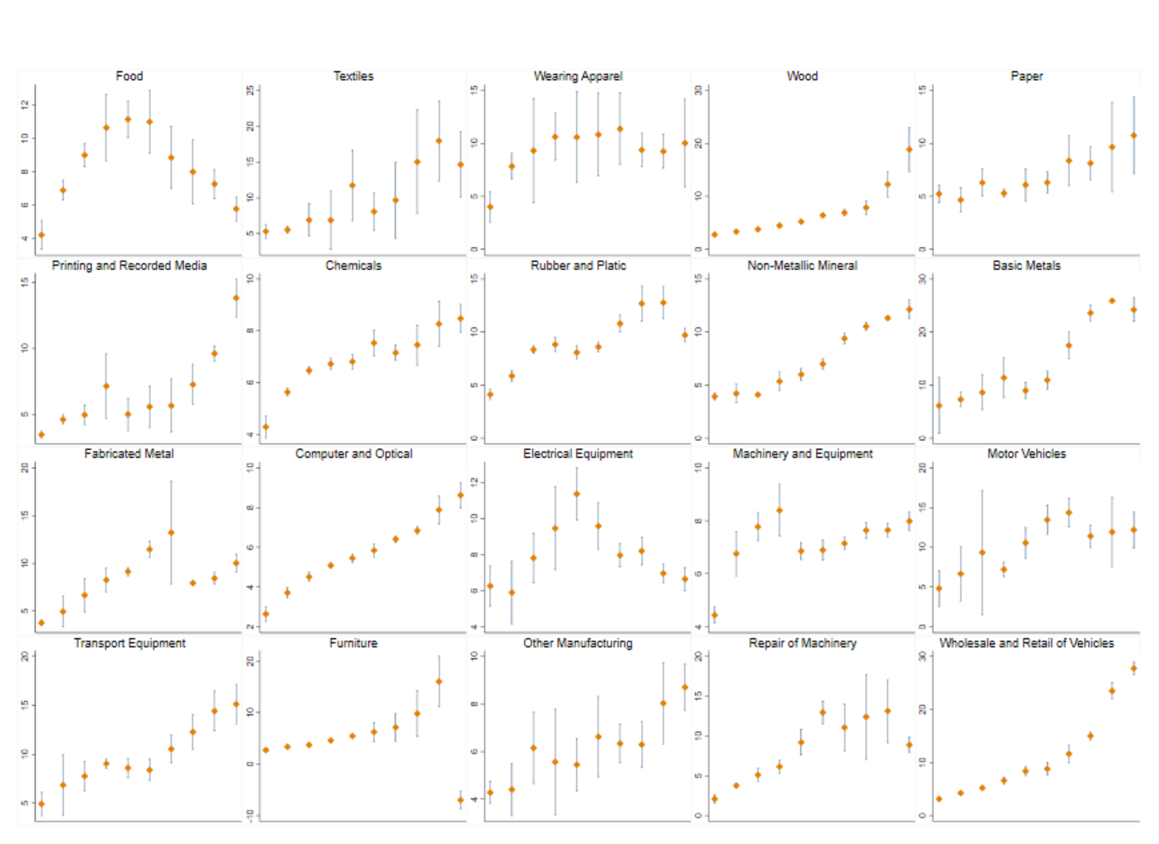


Figure 2: Results are ordered from the lowest to the highest sales decile as we move from left to right. Diamonds indicate coefficient estimates of the median output response and lines indicate 95% confidence intervals obtained by bootstrapping using data for 2010. Pulling observations across years is 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

All Manufacturing Industries

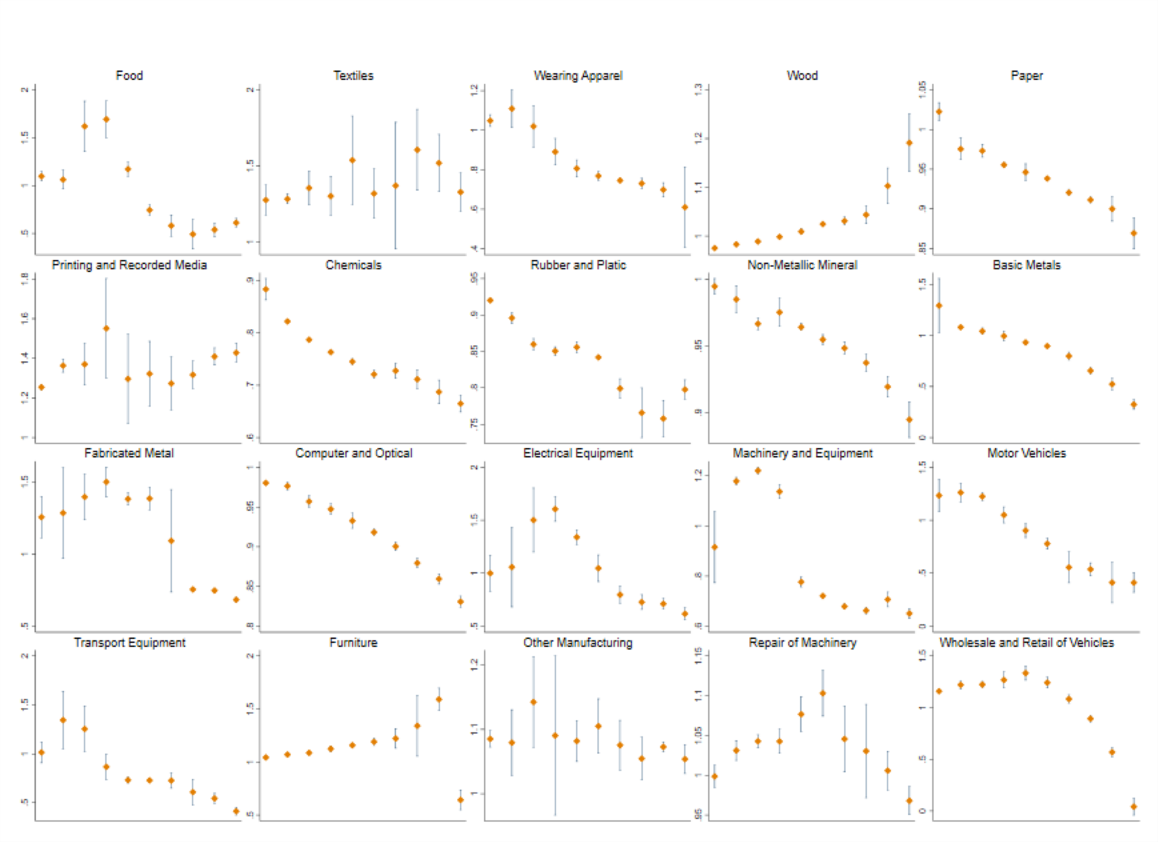


Figure 3: Results are ordered from the lowest to the highest sales decile as we move from left to right. Diamonds indicate coefficient estimates of the median price pass-through and lines indicate 95% confidence intervals obtained by bootstrapping using data for 2010. Pulling observations across years 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.

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 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. Heterogeneity in firm markups and pass-through as given by equation (12) implies that the re-allocation channel will be non-zero. However, because a profit tax does not affect the output decision of the firm, 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[20], I consider the welfare implications of changing its schedule in a differential way.[21] 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 in Section 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 not a function of markup only but also depends on the output response of the firm. Intuitively, the planner cares about the amplitude of a unit firm-specific tax shock, which is determined by the endogenous response of firm output. Having defined the welfare weights, I then check their empirical distribution in the data and leverage the near monotonicity with respect to firm sales to examine a two-tier reform of the value added tax.

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

$$(1 - t(c))\lambda(xu''(x) + u'(x)) = cv'(x) \tag{27}$$

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 so that the perturbed tax schedule is given by $t(c) + \theta\hat{t}(c)$, where $\theta \in \mathbb{R}$ parametrizes the size of the reform.

[20] 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 being in 2011 and took the standard rate from 17.5% to 20%.

[21] Differential payroll taxation by firm size could also be studied in this framework but requires more realistic assumptions on the labour market which lays outside the scope of this paper.

While in the benchmark case, we considered a cost-shock that shifted 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 slope of the tangent lines of the two curves at the initial equilibrium so we get that

$$\frac{\hat{x}(c)}{x(c)} = [\epsilon_{mr}(x) + \epsilon_{mc}(x)]^{-1} \left(\frac{\hat{\lambda}}{\lambda} - \frac{\hat{t}(c)}{1 - t(c)} \right) \tag{28}$$

The incidence on the demand index is given by

$$\frac{\hat{\lambda}}{\lambda} = \int \frac{\hat{t}(c)}{1 - t(c)} \tilde{s}(c) \; dc \quad \text{where} \quad \tilde{s}(c) = \frac{(1 - t(c))s(c)g(c)}{\int (1 - t(c))s(c)g(c) \, dc} \tag{29}$$

Similarly to equation 14, the demand index response is given by the average tax shock change. The distinction is that it is the sales shares that are used as weights rather than the variable cost shares. Furthermore, when constructing the shares we need to use the post-tax sales of each firm. If in the initial equilibrium, the tax rate is the same for all firms there will be no distinction between these two measures of sales shares.

5.1.1 Effect on Government Revenue

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

$$\hat{\mathcal{R}}(\hat{t}) = \int \hat{t}(c)s(c) \; dG(c) + \int t(c)\hat{s}(c) \; dG(c) \tag{30}$$

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 (28) to rewrite the tax incidence formula in terms of firm-level elasticities.

$$\hat{\mathcal{R}}(\hat{t}) = \int \hat{t}(c)s(c) \; dG(c) + \int t(c)s(c) \frac{\Delta(c)}{\mu_f(c)} \frac{\hat{t}(c)}{1 - t(c)} \; dG(c) + \frac{\hat{\lambda}}{\lambda} \int t(c)s(c) \left[1 + \frac{\Delta(c)}{\mu_f(c)} \right] \; dG(c) \tag{31}$$

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.

5.1.2 Effect on Aggregate Utility

From the decomposition in equation 19, we see that the selection effect depends on the utility of the least productive variety. Because I never solve for the underlying utility function I cannot identify $u(x_d)$ and instead I drop that channel from my empirical analysis. Note that because the selection effect scales in the sales share of the cut-off firm and in the data there is very large concentration of sales in top firms, this effect is likely to be small. Following a tax reform, the impact on aggregate utility is

$$\lambda \hat{U} = -\mathcal{M}\hat{\mathcal{R}} + \int \left(1 - \frac{\mathcal{M}}{\mu_f}\right) \frac{\hat{x}}{x} \tilde{s}(c)\, dc \tag{32}$$

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 aggregate markup. The second term is the usual reallocation channel, with firm-level output responses determined as in equation 28.[22]

5.1.3 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. I have written the tax perturbation so that θ is the shock to the initial retention rate.

Total Welfare Impact

Because there is homogeneity in taste and income across consumers I assume that the government evaluates social welfare just as the representative consumer does.[23] Let ψ be the marginal value of public 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

[22] Note that the terms in this decomposition correspond to the first and last term in equation 19 which we derived from a cost-push shock. There is also a selection effect following a tax reform which I have excluded from here because I do not estimate in the data.

[23] 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. Goods that are considered necessities like food are zero-rated while domestic heating fuel is taxed at a reduced 5% rate. These types of considerations have been studied previously in papers such as cite. They are tangential to the issue of using taxes to improve the allocative efficiency of the market equilibrium and hence are better understood separately.

government revenue $\widehat{\mathcal{W}} = \lambda \hat{U} + \psi \hat{\mathcal{R}}$. For the elementary reform defined in the previous section, this expression evaluate to

$$\widehat{\mathcal{W}} = \tilde{s}(c^*) \Big\{ -(\omega(c^*) - \bar{\omega}) + (\psi - \mathcal{M})\hat{\mathcal{R}}(c^*) \Big\} \tag{33}$$

where

$$\omega(c) = \left(1 - \frac{\mathcal{M}}{\mu_f(c)}\right) \Delta(c) \quad \text{and} \quad \bar{\omega} = \int \omega(c)\tilde{s}(c)\, dc \tag{34}$$

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.[24] Equation 33 tells us 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, $\omega(c)$ is zero for all firms. Unsurprisingly, there cannot be welfare improving sales tax reforms if the economy is already on the Pareto frontier. For any other demand system, 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 $u(\cdot)$ but also depend crucially on the supply-side since it is the distribution of firm productivities that determines the demand index and the aggregate markup in equilibrium.

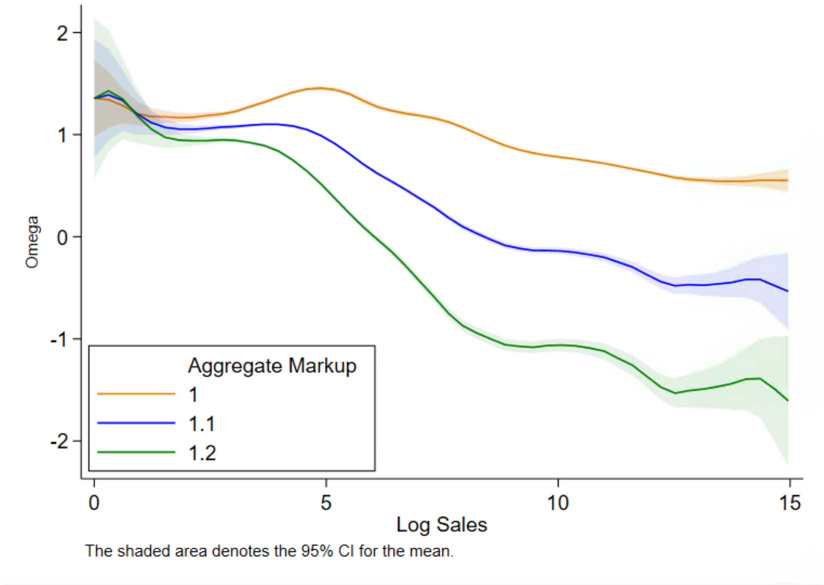
Empirical Properties of ω

We are interested in how the welfare weights $\omega(c)$ change 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.[25] In the data, firm markups typically fall with size while

[24] We could generalize this model to allow for the provision of public goods which together with the existence of constraints on taxation levels could lead to the case where $\psi > \lambda U$. The question addressed here is not whether the private sector is taxed too little (or too much) but whether we can improve on welfare given our current level of taxation.

[25] One can think of this in terms of a partial equilibrium demand supply diagram where the first term measure the distance between the mr and mc curve while the second term tells us the horizontal change in output as a result of shocking one of the curves.

the output response increases. These two forces push in opposite directions and therefore the slope of ω will in general depend on our calibration of the aggregate markup. Figure 4 plots the mean welfare weight for all firms in the sample for three different values of $\mathcal{M}$.



The shaded area denotes the 95% CI for the mean.

Figure 4: 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 aggregate markup. Intuitively, a higher $\mathcal{M}$ puts more weight on the slope of the markup-size relationship which therefore leads to a steeper decline with size. Most importantly though, the slope is still negative even for the lower bound of $\mathcal{M}$.

In Figure 5 I plot welfare weight-firm size relationship for each of the six sectors separately to check for any important sectorial differences.

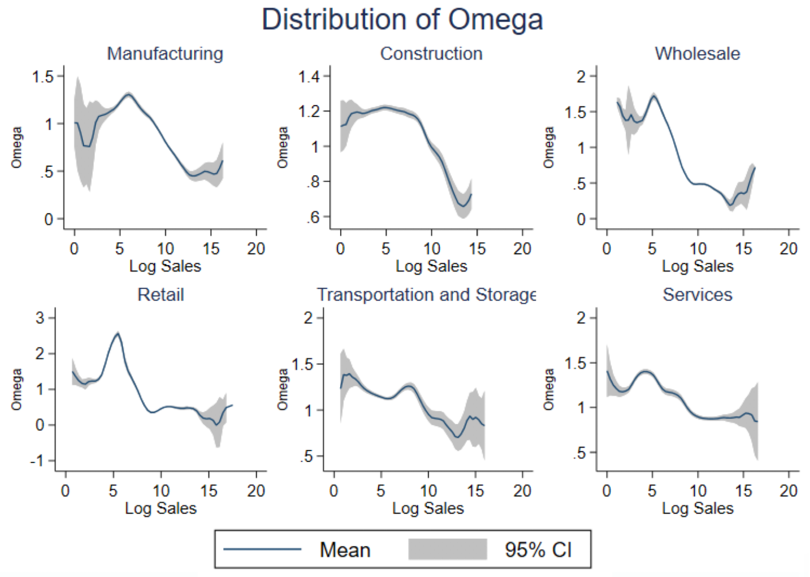


Figure 5: Average welfare weights by firm size for each Sector in 2010.

What is striking is that the monotonicity of this relationships holds up pretty well even when we split firms by sectors. For the smallest of firms we tend to have a dip before the welfare weight increases again but because these firms collectively account for a very small proportion of sales share, variations in ω in this part of the distribution will be much less important.[26]

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 (also equivalent to the sales tax shock since it's a linear tax) for firms below and above the cost-level c^* respectively. I also impose that the tax change be *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

[26] This follows immediately from Equation 33 which shows that the welfare effect of changing the tax rate of any particular firm scales with the sales share of that firm type.

firms respectively, which add up to 1 by definition. Using formula 31 to impose that the tax incidence of the bracket tax reform be exactly zero we derive the following expression for the ratio of the tax shocks.

$$\frac{\theta_2}{\theta_1} = -\frac{S^1}{S^2}\frac{1+t(\bar{\Delta}_s - \bar{\Delta}_s^1)}{1+t(\bar{\Delta}_s - \bar{\Delta}_s^2)} \tag{35}$$

where

$$\bar{\Delta}_s^1 = \int_0^{c^*} \Delta_s(c)\tilde{s}(c)\,dc \quad \text{and} \quad \bar{\Delta}_s^2 = \int_{c^*}^{c_d} \Delta_s(c)\tilde{s}(c)\,dc \tag{36}$$

The first term in equation 40 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 33 and aggregating over all the types we get the total welfare effect of the bracket reform determined by $\{\theta_1, \theta_2, c^*\}$.

$$\lambda\hat{U} = -\theta_1\left\{S^1(\bar{\omega}^1 - \bar{\omega}) + S^2\frac{\theta_2}{\theta_1}(\bar{\omega}^2 - \bar{\omega})\right\} \tag{37}$$

where

$$\bar{\omega}^1(c) = \int_0^{c^*} \omega(c)\tilde{s}(c)\,dc \quad \text{and} \quad \bar{\omega}^2(c) = \int_{c^*}^{c_d} \omega(c)\tilde{s}(c)\,dc \tag{38}$$

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 to the ration of sales for each group and plugging this into equation 37 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. On top of that, the sales share of the reference group (which in this case is the high productivity firms) also shows up because it determines the size of the transfer and hence the how big the impact on the economy is.

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 industries will not be homogenous across industries as the distribution of sales varies significantly by industry. Weak separability of preferences implies that we can solve for the demand index response of each sector separately as given by equation 29. Applying this to our bracket reform we get that

$$\frac{\hat{\lambda}^j}{\lambda^j} = \theta^1 S^{1j} + \theta^2 S^{2j} \tag{39}$$

This in turn allows us to extend the formulae for the *revenue neutral* ratio of tax shocks when we move to the multiple sector case.

$$\frac{\theta_2}{\theta_1} = -\frac{\sum_j \alpha^j S^{1j}(1 + t[\bar{\Delta}_s^j - \bar{\Delta}_s^{1j}])}{\sum_j \alpha^j S^{2j}(1 + t[\bar{\Delta}_s^j - \bar{\Delta}_s^{2j}])} \tag{40}$$

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. Leveraging our result from section 2.5, we have that the welfare statistic is just the sum of the industry-specific welfare impact given in equation 37 and weighted by the industry sales shares α^j. Finally, to make results comparable across different calibrations of the aggregate markup I translate the welfare measure from the money metric one to a *percentage utility change*.[27] The results from this policy experiment are shown in Figure 6.

[27] The money metric measure is not directly comparable for economies with different aggregate markup because it is the aggregate markup that determines the price level in this economy. In order to convert our measure in equation 37 to utils we simply need to divide by $\mathcal{M}$.

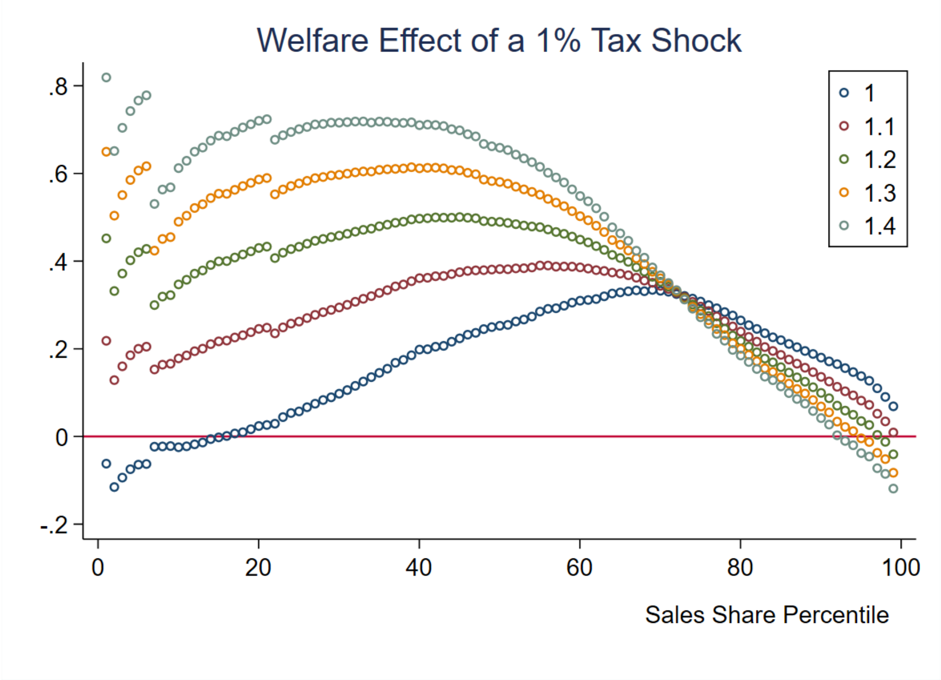


Figure 6: Welfare effect of a 1% tax increase on firms larger than the given sales percentile.

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 aggregate markup $\mathcal{M}$ is. A higher aggregate markup 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. Higher $\mathcal{M}$ also implies that the maximum is reached for lower sales threshold.

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 $\theta_1 = 0.05$.[28] 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 £2m in 2010. Although this tax reform is far from eliminating all markup distortions in the market equilibrium it achieves a pretty large welfare gain.

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

6 Conclusion

This paper provides an analytical formula for the welfare effects of general shocks in a model of monopolistic competition, heterogeneous firms and variable markups. I decompose this effect into three channels: a direct, selection and a reallocation channel. The latter depends not only on firm markups but also on the firm-level output response.

In the empirical application, I allow for many industries and an unrestricted pattern of substitution within each industry. To recover the demand-side elasticities I exploit the commonly used assumption of production function homogeneity. This allows me to estimate the output response non-parametrically from the observed distribution of sales and variable costs in the cross-section of firms. I apply this method to a UK dataset and find that for almost all industries markups decrease with firm size while the output response 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 VAT tax change that increases the VAT rate from 20% to 24% for firms with sales larger than £2m and 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 aggregate markup and support tax relief for small and medium firms at the expense of large ones.

References

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

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

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.

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

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

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.

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*.

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

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., & 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.

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.

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.

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

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

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.

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.

A Model Derivations

A.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 c and simply use x instead. Taking a first order approximation to $\lambda(xu''(x) + u'(x)) = cv'(x)$ we get

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

The first terms on both sides will cancel because they just give the initial equilibrium ($MR = MC$) condition and using our definition of ϵ and ρ we can show that 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

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

A.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

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

where we have used the fact that the cut-off firm must be making exactly zero profits. Similarly, if there is no selection of firms in the initial equilibrium then there would be no effect on expected profit for a small perturbation.

We 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 (aa) multiply both integrals by M_e and apply the definitions of total variable cost 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.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

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

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.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 shown below.

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

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 19.

A.5 Supply Side

Under Assumption 3 we can write the generalized cost-minimization problem of the firm as

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

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial M} &= -p^M + \psi\omega F_M = 0 \\
\frac{\partial \mathcal{L}}{\partial L} &= -p^L + \psi\omega F_L = 0
\end{aligned}$$

Now, we want to show that the ration $\frac{L}{M}$ is independent of the total output Y.

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

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

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

$$Y = \omega F\left(M^*1, M^*\frac{L^*}{M^*}, K\right) = \omega(M^*)^r F\left(1, \gamma, K\right)$$

This allows us to solve for the amount of materials that the firm will purchase

$$M^* = \left(\frac{Y}{\omega F(1, \gamma(p^M, p^L, K), K)}\right)^{1/r}$$

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

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

A.6 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 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

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