Journal of Monetary Economics 121 (2021) 1–14

Contents lists available at ScienceDirect

Journal of Monetary Economics

journal homepage: www.elsevier.com/locate/jmoneco

Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data[☆]

Steve Bond[a], Arshia Hashemi[b], Greg Kaplan[b,c,*], Piotr Zoch[d,e]

[a] University of Oxford, Oxford, UK
[b] The University of Chicago, Chicago, USA
[c] National Bureau of Economic Research, Cambridge, USA
[d] University of Warsaw, Warsaw, Poland
[e] FAME|GRAPE, Warsaw, Poland

ARTICLE INFO

Article history:
Received 2 May 2021
Accepted 3 May 2021
Available online 21 May 2021

JEL classification:
D2
D4
L1
L2

Keywords:
Markups
Output elasticity
Revenue elasticity
Production functions

ABSTRACT

The ratio estimator of the markup is the ratio of the output elasticity for a flexible input to that input's cost share in total revenue. We highlight identification and estimation issues pertaining to this ratio estimator, when firm-level output prices are not observed. If the revenue elasticity for a flexible input is used in place of the output elasticity, then profit maximization implies that the ratio estimator is identically equal to one, and thus is uninformative about markups. Concerning estimation of output elasticities: with only revenue data, profit maximization also implies that the output elasticity is not identified non-parametrically from estimation of the revenue production function, if firms have market power. Even with separate output price and quantity data, it is challenging to estimate the output elasticity consistently if there are non-linear productivity dynamics and firms face heterogeneous demand schedules, with unobserved variation in a demand shifter.

1. Introduction

The production approach to markup estimation identifies a firm's markup as the ratio of the output elasticity for a flexible input to that input's cost share in total revenue. We refer to this estimator of the markup as the *ratio estimator.*[1] This paper evaluates the usefulness of the ratio estimator of the markup in settings in which the empirical measure of output is revenue, rather than physical quantity, and firms have market power in output markets.

The production approach was pioneered by Hall (1986, 1988), in his estimates of aggregate industry-level markups. The recent literature extends the Hall methodology to estimate microeconomic firm- or establishment-level markups (see De Loecker and Warzynski (2012), De Loecker et al. (2020), and many others). The microeconomic ratio estimator is widely

☆ We thank the editors, Laurence Ales and Ariel Zetlin-Jones, and our discussant Jan De Loecker. We are grateful to Dan Ackerberg, Klaus Adam, Susanto Basu, Ali Hortaçsu, Oleg Itskhoki, Jordi Jaumandreu, Peter Neary, Måns Söderbom, Chad Syverson, and conference participants at the 2020 NBER Summer Institute IO and the November 2020 CRNYU Conference on Public Policy (Market Power and Macro Policy) for helpful comments and suggestions. Tuğçe Türk provided excellent research assistance.

* Corresponding author at: Department of Economics, The University of Chicago, 1126 East 59th Street, Chicago, IL, 60637, USA.
E-mail address: gkaplan@uchicago.edu (G. Kaplan).

[1] Strictly speaking, this is the estimand of the ratio estimator, since the output elasticity is not measured directly and typically has to be estimated.

https://doi.org/10.1016/j.jmoneco.2021.05.004
0304-3932/

used across the IO, trade, and macro literatures, and also serves as a popular tool for characterizing the distribution of markups in several economic models, including granular business cycles (Burstein et al., 2020), misallocation in production networks (Baqaee and Farhi, 2020), and monopolistic competition with heterogeneous firms (Mrázová et al., 2018). Many important pitfalls of the ratio estimator have already been discussed (see Traina (2018), Basu (2019), and Syverson (2019)).[2] The issues that we raise in this paper should serve as a caution against drawing inferences from firm-level markup estimates based on the production approach in settings in which firm-level output prices are unobserved.

When physical output quantities are unobserved, as is the case in most of the papers cited above, it is common practice to proxy output with revenues or value added, deflated with common industry-level price deflators. This approach uses the *revenue* elasticity for a flexible input, in place of the *output* elasticity, in the numerator of the ratio estimator. Klette and Griliches (1996) show that, if firm-level output prices are unobserved and correlated with firms' input choices, then estimators of the revenue elasticity are downward-biased estimators of the output elasticity. We show that the implications of this so-called omitted price bias for identifying markups are much more severe than just generating downward bias in the ratio estimator. Under the standard assumption that the flexible input and the output price are determined from a static profit maximization problem, the ratio estimator that uses the revenue elasticity in place of the output elasticity is identically equal to one, and therefore contains no useful information about markups. We pursue the implications of this observation for the identification and estimation of markups using the ratio estimator of the production approach.

The first part of our paper concerns the identification of markups using the ratio estimator. In Section 2.1, we abstract from estimation issues and suppose that the revenue elasticity and output elasticity are known. We then assess the implications for identifying markups of using one elasticity versus the other in the numerator of the ratio estimator. The main takeaway from this section is that it is essential that the output elasticity, rather than the revenue elasticity, is used in the numerator of the ratio estimator. Even in this best case scenario in which population elasticities are known, replacing the output elasticity with the revenue elasticity removes all information about the markup from the ratio estimator. This result follows from imposing firms' static profit maximization conditions in addition to cost minimization.

In Section 2.2, we raise two additional challenges for identifying markups that arise even when the output elasticity is used in the numerator of the ratio estimator. First, we show that if the input that is used to construct the ratio estimator incurs costs of adjustment, then the ratio estimator reflects the shadow cost of adjusting the input as well as the markup. Second, we show that if the input that is used to construct the ratio estimator is used by firms both to produce output and to influence demand, then the ratio estimator generates a downward-biased estimate of the markup. Such inputs include labor and materials used for marketing, product design, or other sales-related purposes (see Syverson (2011) for a related discussion in the context of productivity estimation).

The second part of our paper concerns the estimation of the output elasticity that is required to identify markups using the ratio estimator. In Section 3.1, we show that in the usual setting in which the researcher observes only revenue, and does not have separate information on the price and quantity of output, the output elasticity for a flexible input is not identified non-parametrically from estimation of the revenue production function, if the flexible input and the output price are determined from a static profit maximization problem. There exist parametric restrictions on the forms of the quantity production function and the inverse demand curve under which the output elasticity for a flexible input may be estimated consistently at one point in the parameter space, but these special cases appear to be of limited empirical relevance. The main takeaway from this section is that firm-level data on output prices is needed to obtain credible estimates of the output elasticity for a flexible input from the estimation of a production function when firms have market power.

We also show that even with firm-level data on output prices, it is still challenging to obtain consistent estimates of output elasticities for flexible inputs, particularly if there are non-linear dynamics in total factor productivity. With such non-linearity, the estimators that are widely used in this context do not estimate the output elasticity for a flexible input consistently if firms face heterogeneous demand curves with *unobserved* variation across firms in a demand shifter, or if only a firm-level price *index* is available. In Section 3.2 we briefly discuss the problem of estimating revenue elasticities; standard estimators do not estimate revenue elasticities consistently in panels with many firms and few time periods, if there is unobserved heterogeneity across firms in markups.

Overall, the identification and estimation issues that we highlight cast serious doubt over whether anything useful can be learned about trends or heterogeneity in markups from applying the ratio estimator in settings in which output prices and quantities are unobserved.

2. Difficulties in identifying markups from production function elasticities

In this section, we clarify the conditions under which markups can be identified from knowledge of production function elasticities and the cost shares of flexible inputs in total revenue.

[2] De Loecker and Goldberg (2014) also discuss the implications of unobserved output and input price heterogeneity across firms for production function estimation in settings with limited heterogeneity in markups (e.g. monopolistic competition with constant elasticity of substitution demand) and discuss partial solutions in these cases. We extend this line of research by studying the implications of omitted output prices for (i) identifying markups using the ratio estimator under general demand conditions and market structures and (ii) estimating output elasticities in settings with imperfect competition and unobserved heterogeneity in demand.

In Section 2.1, we emphasize that knowledge of the *output* elasticity for a flexible input, as opposed to knowledge of the *revenue* elasticity for that input, is essential in this regard. In Section 2.2, we highlight two key assumptions that are required to identify markups. Throughout this section, we abstract from firm heterogeneity in productivity, demand, and input prices; we consider these features in Section 3, where we discuss challenges to estimating the population elasticities that are treated as known in this section.

2.1. Output elasticities versus revenue elasticities

We begin by describing the cost minimization problem upon which the production approach to markup estimation is based. Each firm i in period t produces output Q_{it} using a production technology with J flexible inputs:

$$Q_{it} = \mathcal{F}\left(X_{it}^1, \ldots, X_{it}^J\right).$$

The only restriction that we place on the production function $\mathcal{F} : \mathbb{R}_+^J \to \mathbb{R}_+$ is that it is twice continuously differentiable in each of the inputs $\left(X_{it}^1, \ldots, X_{it}^J\right)$.[3] Denote by $\boldsymbol{W}_t := \left(W_t^1, \ldots, W_t^J\right)$ the corresponding vector of input prices, over which firms have no influence and take as given. The output elasticity of input X_{it}^j is defined as

$$\theta_{it}^{Q,j} := \frac{X_{it}^j}{Q_{it}} \frac{\partial \mathcal{F}(\cdot)}{\partial X_{it}^j}.$$

We define the estimand of the ratio estimator of the markup using the *output elasticity* $\theta_{it}^{Q,j}$ in the numerator as

$$\mu_{it}^{Q,j} := \frac{\theta_{it}^{Q,j}}{\alpha_{it}^j}$$

where $\alpha_{it}^j := \left(W_t^j X_{it}^j\right)/(P_{it} Q_{it})$ denotes the cost share of input X_{it}^j in total revenue $R_{it} := P_{it} Q_{it}$.

The firm's static cost minimization problem involves choosing its inputs $\left\{X_{it}^j\right\}_{j=1}^J$ to minimize total (variable) cost subject to producing a target level of output Q_{it}

$$\begin{aligned} \mathcal{C}(Q_{it}; \boldsymbol{W}_t) \; := \; & \min_{\{X_{it}^j\}_{j=1}^J} \left\{ \sum_{j=1}^J W_t^j X_{it}^j \right\} \\ & \text{s.t. } \; \mathcal{F}\left(X_{it}^1, \ldots, X_{it}^J\right) \geq Q_{it}. \end{aligned}$$

The total cost function $\mathcal{C}(Q_{it}; \boldsymbol{W}_t)$ is the solution to this cost minimization problem. Let λ_{it} denote the Lagrange multiplier on the output constraint. The first order condition with respect to the input X_{it}^j is

$$W_t^j = \lambda_{it} \frac{\partial \mathcal{F}(\cdot)}{\partial X_{it}^j}. \tag{1}$$

By the envelope theorem, the Lagrange multiplier, which measures the shadow value of relaxing the output constraint, equals marginal cost

$$\lambda_{it} = \frac{\partial \mathcal{C}(\cdot)}{\partial Q_{it}}.$$

The markup $\mu_{it} := P_{it}/\lambda_{it}$ is defined as the ratio of the output price P_{it} to marginal cost λ_{it}. Invoking the envelope theorem, multiplying both sides of the cost minimization first order condition (1) by X_{it}^j, dividing both sides by $P_{it} Q_{it}$, and rearranging yields

$$\mu_{it}^{Q,j} = \mu_{it}$$

for each flexible input X_{it}^j. That is, the estimand of the ratio estimator using the output elasticity is equal to the markup, as shown by De Loecker and Warzynski (2012).

Suppose now that the researcher does not have knowledge of the output elasticity $\theta_{it}^{Q,j}$, but rather only has knowledge of the revenue elasticity, defined as

$$\theta_{it}^{R,j} := \frac{X_{it}^j}{R_{it}} \frac{\partial R_{it}}{\partial X_{it}^j}.$$

[3] For simplicity, we treat all inputs $\left\{X_{it}^j\right\}_{j=1}^J$ as fully flexible, but this is not essential to the points we make in this section. If a subset of the inputs were fully fixed or predetermined, we could work with the conditional cost function. In Appendix A.2, we show that our main results are robust to a subset of the inputs being partially fixed and subject to adjustment costs.

We define the estimand of the analogous ratio estimator of the markup using the *revenue elasticity* $\theta_{it}^{R,j}$ in the numerator as

$$\mu_{it}^{R,j} := \frac{\theta_{it}^{R,j}}{\alpha_{it}^{j}} .$$

We assess whether the ratio estimator using the revenue elasticity $\theta_{it}^{R,j}$ in place of the output elasticity $\theta_{it}^{Q,j}$ is informative about the markup. In this section, we focus on monopolistic competition to illustrate our main theoretical result in the simplest of market structures with imperfect competition. In Appendix A.1, we show that our result is robust to both Bertrand and Cournot forms of oligopolistic competition with strategic interactions between firms.

A monopolistic firm faces an arbitrary inverse demand schedule:

$$P_{it} = \mathcal{P}(Q_{it}).$$

The only restrictions that we impose on the function $\mathcal{P} : \mathbb{R}_+ \to \mathbb{R}_+$ are that it is twice continuously differentiable and non-increasing in Q_{it}. The absolute value of the price elasticity of demand is defined as

$$\eta_{it} := \left| \frac{P_{it}}{Q_{it}} \frac{\partial Q_{it}}{\partial P_{it}} \right| > 1 .$$

We can explicitly write the revenue elasticity $\theta_{it}^{R,j}$ in terms of the demand elasticity η_{it} and the output elasticity $\theta_{it}^{Q,j}$

$$\theta_{it}^{R,j} = \left(\frac{\eta_{it} - 1}{\eta_{it}} \right) \theta_{it}^{Q,j} . \tag{2}$$

Monopolistic firms with market power in output markets face a finite demand elasticity $\eta_{it} < \infty$. It is then apparent from Eq.(2) that the revenue elasticity $\theta_{it}^{R,j}$ is strictly less than the output elasticity $\theta_{it}^{Q,j}$.[4]

Taking the total cost function $\mathcal{C}(Q_{it}; \boldsymbol{W}_t)$ from cost minimization as given, the static profit maximization problem involves choosing the output quantity Q_{it} to maximize profits subject to the demand schedule

$$\begin{aligned} &\max_{Q_{it}} \{P_{it} Q_{it} - \mathcal{C}(Q_{it}; \boldsymbol{W}_t)\} \\ \text{s.t.} \quad & P_{it} = \mathcal{P}(Q_{it}). \end{aligned}$$

The first order condition from profit maximization recovers the markup μ_{it} as a function of the demand elasticity η_{it}

$$\mu_{it} = \frac{\eta_{it}}{\eta_{it} - 1} . \tag{3}$$

Imposing both cost minimization and profit maximization, we obtain

$$\begin{aligned} \mu_{it}^{R,j} &= \frac{\theta_{it}^{R,j}}{\alpha_{it}^{j}} \\ &= \left(\frac{\eta_{it} - 1}{\eta_{it}} \right) \frac{\theta_{it}^{Q,j}}{\alpha_{it}^{j}} \\ &= \frac{1}{\mu_{it}} \mu_{it} \\ &= 1 . \end{aligned}$$

The first equality follows from the definition of $\mu_{it}^{R,j}$, the second from Eq. (2), and the third from the first order conditions from cost minimization (1) and profit maximization (3). It is apparent that the ratio estimator using the revenue elasticity contains no useful information about the markup, except in the very special case of perfect competition under which the markup equals one.[5]

The result $\mu_{it}^{R,j} = 1$ is a consequence of profit maximization and, importantly, does not depend on the particular details of the profit maximization problem, such as the functional form of the demand schedule or the market structure. To understand why, consider an industry with N competing firms indexed by $i \in \{1, \ldots, N\}$. Let $\boldsymbol{Q}_{-it} := \{Q_{kt}\}_{k \neq i}$ denote the vector of quantities of all $(N-1)$ competitors of firm i, and let $\boldsymbol{Q}_t := \left(Q_{it}, \boldsymbol{Q}_{-it}\right)$. Consider an arbitrary inverse demand schedule $P_{it} = \mathcal{P}_i\left(Q_{it}, \boldsymbol{Q}_{-it}\right)$, for $i \in \{1, \ldots, N\}$, constituting a one-to-one mapping between any vector of quantities $\boldsymbol{Q}_t$ and corresponding vector of prices $\boldsymbol{P}_t := (P_{1t}, \ldots, P_{Nt})$.[6] Let $R_{it} = \mathcal{R}_i\left(Q_{it}, \boldsymbol{Q}_{-it}\right) := \mathcal{P}_i\left(Q_{it}, \boldsymbol{Q}_{-it}\right) Q_{it}$ denote the revenue function for firm i in

[4] Intuitively, a monopolistic firm facing a strictly downward-sloping demand schedule must reduce its output quantity to increase its price. Hence, the revenue elasticity is strictly less than the output elasticity.

[5] Under perfect competition, the revenue elasticity equals the output elasticity because firms have no influence over output prices.

[6] We rule out demand schedules for which a flexible input X_{it}^{j} enters as an additional argument of the function $\mathcal{P}_i(\cdot)$. We study the implications of such cases for identifying markups in Section 2.2.

period t. This formulation allows for a range of market structures, including monopolistic competition (as $N \to \infty$) and both Bertrand and Cournot forms of oligopolistic competition (for a finite N).

Imposing only cost minimization, we have

$$\begin{aligned}
\mu_{it}^{R,j} &= \frac{\theta_{it}^{R,j}}{\alpha_{it}^{j}} \\
&= \frac{X_{it}^{j}}{R_{it}} \frac{\partial \mathcal{R}_i(\cdot)}{\partial X_{it}^{j}} \frac{1}{\alpha_{it}^{j}} \\
&= \frac{d\mathcal{R}_i(\cdot)}{dQ_{it}} \frac{1}{P_{it}} \frac{X_{it}^{j}}{Q_{it}} \frac{\partial \mathcal{F}(\cdot)}{\partial X_{it}^{j}} \frac{1}{\alpha_{it}^{j}} \\
&= \frac{d\mathcal{R}_i(\cdot)}{dQ_{it}} \frac{1}{P_{it}} \frac{\theta_{it}^{Q,j}}{\alpha_{it}^{j}} \\
&= \frac{d\mathcal{R}_i(\cdot)}{dQ_{it}} \frac{\mu_{it}}{P_{it}} \\
&= \frac{d\mathcal{R}_i(\cdot)}{dQ_{it}} \left[\frac{\partial \mathcal{C}(\cdot)}{\partial Q_{it}} \right]^{-1}
\end{aligned}$$

for each flexible input X_{it}^{j}. The first equality follows from the definition of $\mu_{it}^{R,j}$, the second from the definition of $\theta_{it}^{R,j}$, the third from the chain rule and the definition of $R_{it} = P_{it} Q_{it}$, the fourth from the definition of $\theta_{it}^{Q,j}$, the fifth from cost minimization, and the sixth from the definition of the markup μ_{it}. Under cost minimization, but not profit maximization, the estimand $\mu_{it}^{R,j}$ is equal to the ratio of marginal revenue $\frac{d\mathcal{R}_i(\cdot)}{dQ_{it}}$ to marginal cost $\frac{\partial \mathcal{C}(\cdot)}{\partial Q_{it}}$. Additionally imposing profit maximization $\frac{d\mathcal{R}_i(\cdot)}{dQ_{it}} = \frac{\partial \mathcal{C}(\cdot)}{\partial Q_{it}}$ then gives $\mu_{it}^{R} = 1$, for all values of the true underlying markup μ_{it}.

Discussion

For profit maximizing firms, we have established that the estimand of the ratio estimator using the revenue elasticity for a flexible input, in place of the output elasticity, contains no information about the markup. Intuitively, the output elasticity and the revenue elasticity are only equal when a firm is not able to influence its output price by varying its output quantity. But the ability to affect price by changing quantity is the reason why firms with market power charge markups above one.

Our contribution is closely related to Klette and Griliches (1996), who showed that using revenue in place of output to estimate an output elasticity results in a downward bias when firms have market power. In our simple example, this effect is readily seen from Eq. (2), together with the typical assumption that demand curves slope downward. Since the ratio estimator should use the output elasticity in the numerator, Klette and Griliches (1996) is often cited as a reason why using revenue elasticities instead of output elasticities leads to downward-biased estimates of the markup (see for example De Loecker and Warzynski (2012), Section VI). While this is true in a technical sense if the true markup is above one, our result shows that the problem is more fundamental. The bias in the ratio estimator from using the revenue elasticity, in place of the output elasticity, removes all the information about the markup, so that the biased estimator is not informative about the markup at all.

Unfortunately, output quantities Q_{it} are rarely observed for individual firms. Instead, researchers typically only have access to measures of revenues R_{it}. As we explain in Section 3, when firms have market power, it is not possible to learn about the output elasticity $\theta_{it}^{Q,j}$ by estimating a production function specification that uses revenue as the dependent variable, under any reasonable assumptions (and it is challenging even with data on output quantities). With only data on revenues, it is not clear that we can learn anything about the level of markups using the ratio estimator.

Finally, it is useful to bear in mind that if it were somehow possible to recover the output elasticity from knowledge of the revenue elasticity, then it would not be necessary to use the ratio estimator to learn about markups. Under monopolistic competition, one could simply estimate both the output elasticity and the revenue elasticity, and note from Eqs. (2) and (3) that the ratio of these two elasticities is an estimator of the markup. This observation is a reminder that the problem with revenue elasticities that we are highlighting in this section is not one of estimation, but one of identification: any attempt to learn about the output elasticity from the revenue elasticity must implicitly have assumed knowledge of the markup. The resulting output elasticity therefore cannot contain any additional information that is useful in identifying markups.

Since the estimand underlying the ratio estimator is unity when the revenue elasticity is used in the numerator, it is natural to ask why existing empirical work using this approach does not find estimates that are centered around one. In the following section, we mention two additional sources of bias in the ratio estimator that are likely to be reflected in these estimates. In Section 3.2, we explain why even estimates of the revenue elasticity are likely to be biased. Given these sources of bias, it is not surprising that estimates using the ratio estimator obtained with revenue data are not centered around one.

2.2. Two additional difficulties in the interpretation of the ratio estimator

In Section 2.1, we showed that if the revenue elasticity is used in the numerator of the ratio estimator, then the resulting estimand is equal to unity, and contains no information about the markup under the assumption of profit maximization. But when the output elasticity is used in the numerator of the ratio estimator, the resulting estimand correctly recovers the markup under the assumption of cost minimization. In this section, we offer two caveats to this result that apply even in the more favorable case when the output elasticity is known: (i) input adjustment costs, and (ii) inputs that are used not only for production, but also to influence demand.

Input adjustment costs

For the ratio estimator to recover the markup, it is crucial that the input X_{it}^{j} whose output elasticity and cost share are combined is perfectly flexible. Alternatively, as explained in Basu (2019), X_{it}^{j} could be a bundle of inputs, of which at least one component is perfectly flexible, with the other components being fully fixed. However, in reality, inputs rarely fall into one of these two extreme cases. A more realistic intermediate case is to assume that inputs are partially adjustable, in the sense that firms incur costs to adjust their input choices. If the ratio estimator is constructed using an input X_{it}^{j} that is partially adjustable, or using a bundle that contains partially adjustable inputs, then the estimand of the ratio estimator will reflect both the markup and the shadow cost of adjusting those inputs.

To illustrate this point, assume instead that each input X_{it}^{j} is associated with a baseline quantity $\overline{X}_{it}^{j}$ and that the firm incurs adjustment costs when it chooses a quantity of input $X_{it}^{j} \neq \overline{X}_{it}^{j}$. The baseline quantity $\overline{X}_{it}^{j}$ might reflect the input choice from the previous period in a dynamic version of the model. For simplicity, we assume that these costs are given by the smooth convex function $\kappa^{j}\left(X_{it}^{j}\right)$, which satisfies $\kappa^{j}\left(\overline{X}_{it}^{j}\right) = 0$ and $\frac{d\kappa^{j}\left(\overline{X}_{it}^{j}\right)}{dX_{it}^{j}} = 0$. In Appendix A.2 we show that the estimand using the revenue elasticity is then

$$\mu_{it}^{R,j} = \frac{\theta_{it}^{R,j}}{\alpha_{it}^{j}} = 1 + \frac{d\kappa^{j}\left(X_{it}^{j}\right)}{dX_{it}^{j}},$$

and the estimand using the output elasticity is

$$\mu_{it}^{Q,j} = \frac{\theta_{it}^{Q,j}}{\alpha_{it}^{j}} = \mu_{it}\left[1 + \frac{d\kappa^{j}\left(X_{it}^{j}\right)}{dX_{it}^{j}}\right].$$

Thus, even if the output elasticity for an input were known, it is crucial that none of the inputs in the bundle incur adjustment costs, in order for the ratio estimator to recover the markup.[7]

Inputs that influence demand

The framework in Section 2.1 assumed that the inputs X_{it}^{j} are only used to produce output and not also to influence demand. Assume instead that the firm's revenue is given by

$$R_{it} := \mathcal{P}(Q_{it}, D_{it})Q_{it}$$

where D_{it} is an endogenous demand shifter that the firm can influence through the use of inputs according to the function

$$D_{it} = \mathcal{D}\left(X_{it}^{D,1}, \ldots, X_{it}^{D,J}\right)$$

where $X_{it}^{D,j}$ is the amount of input j used in influencing demand, and $X_{it}^{Q,j}$ is the amount of input j used in production. We assume that we can observe only the total quantity of input j used by the firm $X_{it}^{j} = X_{it}^{Q,j} + X_{it}^{D,j}$. In Appendix A.3, we show that the estimand underlying the ratio estimator using the output elasticity then becomes

$$\mu_{it}^{Q,j} = \mu_{it}\left[\frac{\psi_{it}^{Q,j}}{1 + \frac{X_{it}^{D,j}}{X_{it}^{Q,j}}}\right]$$

where $\psi_{it}^{Q,j}$ is the elasticity of $X_{it}^{Q,j}$ with respect to X_{it}^{j} evaluated at the optimum, which shows how an additional unit of X_{it}^{j} is allocated between $X_{it}^{Q,j}$ and $X_{it}^{D,j}$. So if the flexible input is only used for production and not to influence demand (i.e. $\psi_{it}^{Q,j} = 1, X_{it}^{D,j} = 0$), then the ratio estimator using the output elasticity recovers the markup. But if some of the input is used to influence demand, and this component cannot be separated out, then the ratio estimator will be biased. If the firm uses a constant fraction of the input X_{it}^{j} for production, then $\psi_{it}^{Q,j} = 1$ and the ratio estimator is biased downward. For example,

[7] These results assume that observed input costs are $W_t^{j}X_{it}^{j}$ rather than $W_t^{j}X_{it}^{j} + W_t^{j}\kappa^{j}\left(X_{it}^{j}\right)$. If observed input costs also include the adjustment costs, then we would obtain $\mu_{it}^{Q,j} = \mu_{it}\left(\frac{X_{it}^{j} + \frac{d\kappa^{j}\left(X_{it}^{j}\right)}{dX_{it}^{j}}}{X_{it}^{j} + \kappa^{j}\left(X_{it}^{j}\right)}\right)$, which also does not recover the true markup.

if, over time, the input X_{it}^j is increasingly being used to influence demand, then the ratio estimator will fall, without any change in the true markup.

3. Difficulties in estimating production function elasticities when firms have market power

In Section 2, we established that when using the ratio estimator to estimate markups, it is critical to use the *output* elasticity with respect to a flexible input in the numerator, rather than the *revenue* elasticity. In this section, we highlight several difficulties that arise when attempting to estimate the required output elasticity when firms have market power. We also note that it is not straightforward to obtain consistent estimates of the revenue elasticity, particularly if there is unobserved heterogeneity across firms in markups.

3.1. Estimation of the output elasticity for a flexible input

We start in Section 3.1.1 by considering the case in which the researcher observes only revenue, and does not have separate information on the price and quantity of output. We show that in this case the output elasticity for a flexible input is not identified non-parametrically from estimation of the revenue production function. There exist parametric restrictions on the forms of the quantity production function and the inverse demand curve under which the output elasticity for a flexible input may be estimated consistently at one point in the parameter space, but these special cases appear to be of limited empirical relevance for studying heterogeneity in markups.

In Section 3.1.2, we then consider the case in which the researcher observes both revenue and the output price for individual firms, or equivalently has data on output quantities. In this case the output elasticity for a flexible input is identified under reasonable conditions if there is no measurement error in the data on output, or if total factor productivity follows a linear ARMA process. In these cases, output elasticities can be estimated consistently using moment conditions for the quantity production function of the kind suggested by Blundell and Bond (2000).

Even with output quantity data, consistent estimation of the output elasticity for a flexible input is more challenging if output is measured with error and total factor productivity follows a non-linear process. Two stage estimators, of the type suggested by Ackerberg et al. (2015) for the estimation of value added production functions for price-taking firms, have often been used in this context.[8] The measurement error in observed output is eliminated using a first stage regression, which allows non-linear dynamic processes for unobserved total factor productivity to be considered in the second stage. The first stage specification requires a valid control function for total factor productivity, which is obtained by inverting a demand function for the flexible input in which total factor productivity is the *only* unobserved component. This approach cannot be used if the demand curves are firm-specific and there is some unobserved heterogeneity across firms in a demand shifter, as well as in total factor productivity, unless the researcher can also control for variation across firms in marginal costs.[9]

In Section 3.1.3, we consider the case in which the researcher observes both revenue and a firm-specific output price index, but does not have data on output price levels for individual firms. Deflating revenue using the firm-specific output price index results in a measure of output which differs from the true level of output by an unknown multiplicative firm-specific constant, reflecting differences across firms in output prices in the base year. In logarithmic specifications, this measurement error can be accounted for by firm-specific fixed effects, but obtaining consistent estimates of output elasticities then requires these fixed effects to be taken into account. This is also problematic if we need to deal with non-linearity in the dynamic process for total factor productivity. The presence of unobserved firm-specific fixed effects can however be handled if total factor productivity follows a linear ARMA process, using the kind of dynamic panel data estimator for production functions suggested by Blundell and Bond (2000).

3.1.1. Data on revenue

In this section we consider a three factor Hicks-neutral gross output production function for firm i in period t of the form

$$q_{it} = f(k_{it}, l_{it}, m_{it}) + \omega_{it} \tag{4}$$

in which q_{it} is the log of gross output, k_{it}, l_{it} and m_{it} are the logs of observed capital, labor and intermediate inputs respectively, and ω_{it} is the log of total factor productivity, which is observed by the firm but not by the researcher. We treat capital and labor as predetermined inputs, for which the input levels are chosen before the firm has observed ω_{it}.[10] We assume that the level of intermediate inputs is chosen after the firm has observed ω_{it}, and that intermediate inputs do not incur adjustment costs of any kind; that is, we consider intermediate inputs as our example of an input which is flexible in the sense required to construct the ratio estimator of the markup. The object of interest is thus the output elasticity $\theta_{it}^{Q,M} := \partial f(\cdot)/\partial m_{it}$.

[8] See, for example, De Loecker and Warzynski (2012) and De Loecker et al. (2020).

[9] This point has also been made in contemporaneous work by Doraszelski and Jaumandreu (2019).

[10] The predetermined inputs may also be subject to adjustment costs. If so, these adjustment costs do not take the form of foregone production, and do not depend on the level of intermediate inputs in any time period.

The researcher observes neither gross output nor the output price, but only sales revenue or the value of gross output, the log of which is $r_{it} := p_{it} + q_{it}$.[11] To analyze this further, we assume that each firm faces a downward-sloping inverse demand curve of the form

$$p_{it} = p(q_{it}, \xi_{it}) \tag{5}$$

in which ξ_{it} is a demand shifter, which is observed by the firm and may be observed or unobserved by the researcher.

The revenue production function which can be estimated in this setting relates the log of observed revenue to the logs of the observed inputs

$$r_{it} = (p_{it} + q_{it}) = f(k_{it}, l_{it}, m_{it}) + (p_{it} + \omega_{it})\,. \tag{6}$$

The dependence of intermediate inputs (m_{it}) on unobserved total factor productivity (ω_{it}) raises issues for the consistent estimation of the output elasticity $\theta_{it}^{Q,M}$ from the quantity production function (4) that are well known in the context of price-taking firms; we discuss some additional issues which arise when firms have market power in Section 3.1.2 below.

The presence of the output price (p_{it}) in the error term of the revenue production function (6) raises more fundamental issues when firms have market power, and their output price depends on q_{it} from (5), and hence on each of the inputs. This additional source of inconsistency has been analyzed by Klette and Griliches (1996) and termed the 'omitted price bias'.[12] Our contribution here is to show that if the output price and the level of the flexible input are chosen at the same time to maximize the same objective, then the output elasticity $\theta_{it}^{Q,M}$ is not identified non-parametrically from estimation of the revenue production function (6).

The intuition for this result is straightforward in the special case in which all firms face the same inverse demand curve, and we have only common shocks ($\xi_{it} = \xi_t$ for all i) in (5). In this case, with observations on (p_{it}, q_{it}) constrained to lie along this downward-sloping demand curve, any firm-specific shock which increases m_{it} and hence q_{it} also reduces p_{it}. In other words, any informative instrument for m_{it} is correlated with p_{it} and so not a valid instrument in the revenue production function (6). With heterogeneity across firms in the inverse demand curves, the same still applies, except in special cases in which there is no pass through of demand shocks (ξ_{it}) to the output price. In these special cases, if informative proxies for the demand shifter are observed by the researcher, and these are uncorrelated with ω_{it}, then these would provide valid and informative instruments for m_{it} in (6). However, the special cases with zero pass through of demand shocks to the output price require strong parametric restrictions on the form of both the quantity production function (4) and the inverse demand curve (5), such that at best the output elasticity is identified only at one point in the parameter space.

To illustrate this, we assume that the firm chooses its output price (P_{it}) and level of intermediate inputs (M_{it}) to maximize profits in period t, taking the costs of the predetermined inputs as given, or equivalently to maximize revenue net of variable costs

$$\Pi_{it} := P_{it} Q_{it} - P_{it}^M M_{it} \tag{7}$$

subject to the constraints in (4) and (5). Here P_{it}^M is the price of one unit of intermediate inputs for firm i in period t, and p_{it}^M is the log of this price; the input price is observed by the firm, and may be observed or unobserved by the researcher. We assume that the firm takes total factor productivity (ω_{it}), the demand shifter (ξ_{it}), and the flexible input price (p_{it}^M) as given.

The solution equates marginal revenue and marginal variable cost. We can either find the level of intermediate inputs which maximizes net revenue in period t and infer the output price from the inverse demand curve at the resulting level of output, or we can find the output price and quantity which maximize net revenue in period t and infer the required level of intermediate inputs. In either case, we obtain decision rules or policy functions which express both m_{it} and p_{it} as functions of the *same* state variables (k_{it}, l_{it}) and the *same* primitives $(\omega_{it}, \xi_{it}, p_{it}^M)$:

$$\begin{aligned} m_{it} &= m^*(k_{it}, l_{it}, \omega_{it}, \xi_{it}, p_{it}^M) \\ p_{it} &= p^*(k_{it}, l_{it}, \omega_{it}, \xi_{it}, p_{it}^M)\,. \end{aligned} \tag{8}$$

These decision rules then indicate that any informative instrument for m_{it} in (6) will necessarily be correlated with the p_{it} component of the error term, while any instrument that is uncorrelated with p_{it} will not be an informative instrument for m_{it}. Equivalently, if we were able to control adequately for the p_{it} component of the error term in (6) we would have exhausted all the sources of variation in the explanatory variable m_{it}. The explanatory variable m_{it} and the error component p_{it} are 'functionally dependent' in the sense of Ackerberg et al. (2015). Without parametric restrictions, we cannot separately identify the contributions of m_{it} and p_{it} to the log of observed revenue r_{it}.[13]

In this context, variation in the input price p_{it}^M shifts the marginal variable cost schedule; if the demand and marginal revenue schedules are downward-sloping, this variation necessarily also affects the output price. As a result, there are no parametric restrictions that lead to the exclusion of p_{it}^M from the decision rule for the output price in (8). The demand

[11] We abstract here from any difference between sales revenue and the value of production, due to changes in inventories.

[12] See also De Loecker (2011) and De Loecker and Goldberg (2014).

[13] The dependence of the output price on the predetermined inputs also indicates that when firms have market power, we do not have moment conditions of the form $\mathrm{E}[(p_{it} + \omega_{it})|k_{it}, l_{it}] = 0$, versions of which have typically been used in the estimation of revenue production functions.

shocks ξ_{it} shift the marginal revenue schedule, and there are admissible parametric restrictions under which there is zero pass through of the demand shocks to the output price. This would be the case if we have both constant marginal variable cost and the markup does not depend on the level of output.

For example, we may have a Cobb-Douglas gross output production function with increasing returns to scale and a unit output elasticity for the flexible input, and a Constant Elasticity of Substitution (CES) demand curve for each firm.[14] In this case, the demand shocks ξ_{it} affect the level of intermediate inputs but not the output price, and observed proxies for the demand shocks would provide valid and informative instruments for m_{it} in a log-linear version of (6), provided they are also uncorrelated with ω_{it}. This *requires* heterogeneity across firms in the inverse demand curves, and the output elasticity parameter for the flexible input (β_M) is identified *only* at one point ($\beta_M = 1$) in the parameter space. This requirement for the output elasticity to be unity here suggests that these parametric special cases are likely to be of limited empirical relevance. Moreover, since identification here relies on shifts in the demand curve, and shifts in the demand curve would affect the demand for two or more flexible inputs in the same way, the parametric special cases in which this approach could be applied are limited to specifications with a single flexible input, as in the example that we have considered here.[15]

3.1.2. Data on revenue and output price levels

Our result in the previous section indicates that, when firms have market power, data on firm-level output prices is fundamental to obtaining credible estimates of the output elasticity for a flexible input from estimation of a production function. Here we show that even with a quantity measure of output, it is still challenging to estimate this output elasticity consistently, particularly if output is measured with error and total factor productivity follows a non-linear dynamic process.

To simplify the exposition, we now focus on a Cobb-Douglas gross output production function, although the issues we highlight apply for any continuously differentiable gross output production function (see Appendix B.1 for details). We further assume that gross output is measured with a multiplicative error, such that the log of observed output is $y_{it} := q_{it} + \varepsilon_{it}$, where ε_{it} is a mean zero measurement error. The quantity production function to be estimated then has the form

$$y_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + (\omega_{it} + \varepsilon_{it})\,. \tag{9}$$

For simplicity, we choose units such that the mean of ω_{it} is also zero. We assume that the measurement error ε_{it} is uncorrelated with the observed inputs (k_{is}, l_{is}, m_{is}) and with the input price p^M_{is} for any s, t, and is independent across firms.[16] The slope parameters $(\beta_K, \beta_L, \beta_M)$ are the output elasticities, which are assumed to be constant over time and common to all the firms in the sample. Our parameter of interest here is the output elasticity for the flexible input β_M.

We again assume that the firm chooses the level of intermediate inputs to maximize net revenue in (7), subject to the constraints in (9) and (5) and taking $(\omega_{it}, \xi_{it}, p^M_{it})$ as given. Without specifying the form of the inverse demand curve (5), we show in Appendix B.1 that the optimal choice of intermediate inputs satisfies the first order condition

$$m_{it} = \frac{\ln \beta_M}{1-\beta_M} + \left(\frac{\beta_K}{1-\beta_M}\right) k_{it} + \left(\frac{\beta_L}{1-\beta_M}\right) l_{it} + \left(\frac{1}{1-\beta_M}\right)\left(p_{it} - \ln \mu_{it} - p^M_{it} + \omega_{it}\right) \tag{10}$$

where μ_{it} is the markup of price over marginal cost as in Section 2, and we can note that $z_{it} := p_{it} - \ln \mu_{it}$ is the log of marginal cost. The only restriction that we place on the demand curve here is that the output price p_{it} is a weakly decreasing function of gross output q_{it}.

We assume that total factor productivity ω_{it} is independent across firms, and start by considering the special case in which ω_{it} is serially uncorrelated; extensions to more realistic cases in which the unobserved heterogeneity across firms in productivity is persistent over time will be considered below. We consider a setting in which panel data is observed for a large number of firms for a small number of time periods, and asymptotic properties are stated for the case in which the number of firms increases, with the number of time periods treated as fixed.

Under these assumptions, we have the moment conditions $\mathrm{E}[(k_{it}, l_{it})u_{it}] = 0$ where $u_{it} := \omega_{it} + \varepsilon_{it}$ is the error term in (9). If the researcher has data on the input price p^M_{it}, and if these input prices vary across firms in a way that is uncorrelated with ω_{it}, then the price of the flexible input provides a valid and informative instrument for the explanatory variable m_{it} in (9). In that case we have the additional moment condition $\mathrm{E}[p^M_{it} u_{it}] = 0$, and the parameter vector $(\beta_K, \beta_L, \beta_M)$ is identified from the estimation of the quantity production function (9).

[14] That is, we have a gross output production function of the form $q_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + \omega_{it}$ with $\beta_M = \theta^{Q,M}_{it} = 1$ for all i, t, and returns to scale $\nu = \beta_K + \beta_L + 1 > 1$; and an inverse demand curve of the form $p_{it} = \xi_{it} - \eta^{-1} q_{it}$, where $\eta = \eta_{it} > 1$ is the absolute value of the price elasticity of demand for all i, t.

[15] Note that our results in this section apply to a revenue production function which relates revenue to input *quantities*, as in (6). If the specification relates revenue to *expenditures* on (some of) the inputs, so that (some of) the input prices are introduced as additional components of the error term, there may also be parametric special cases in which the output elasticities could be estimated consistently. One example has a constant returns to scale Cobb-Douglas gross output production function, a CES demand curve with the same demand elasticity for all firms, all inputs fully flexible, and all inputs with heterogeneous input prices measured as expenditures.

[16] An alternative interpretation of the two error components in (9) is that ω_{it} denotes the log of the component of total factor productivity that is known by the firm when making input decisions in period t, and ε_{it} denotes the log of an unforecastable productivity shock that is not known by the firm when making input decisions in period t. The presence of the second component (ε_{it}) of the error term here is more important than the particular way we introduce it.

If the researcher does not have data on the price of the flexible input, or if the variation across firms in these input prices is correlated with ω_{it}, the parameter vector $(\beta_K, \beta_L, \beta_M)$ will still be identified here if either: (i) there is variation across firms in the input price p^M_{it} which is persistent over time; or (ii) there is variation across firms in the demand shifter ξ_{it} which is persistent over time and results in persistent variation in the log of marginal cost z_{it}. With persistent variation in either p^M_{it} or z_{it}, the first order condition (10) implies that the lagged input $m_{i,t-1}$ provides a valid and informative instrument for the explanatory variable m_{it} in (9), and in this case we have the additional (informative) moment condition $\mathrm{E}[m_{i,t-1}u_{it}] = 0$.[17]

For price-taking firms, it is well known that identification of the output elasticity for a flexible input from estimation of the quantity production function requires variation across firms in the price of the flexible input.[18] For firms with market power and a single flexible input, persistent variation across firms in demand provides a second mechanism through which the lagged input may be an informative instrument. This could be useful in applications where the researcher has data on expenditure on the flexible input, but does not have firm-level data on the price of the flexible input. Expenditure on the flexible input, deflated using a common price index, provides a suitable measure of the input quantity only if the input price does not vary across firms. This requirement rules out identification of the output elasticity from estimation of the production function for price-taking firms, but may not do so when firms have market power.

We now extend our discussion to consider more realistic cases in which the variation across firms in unobserved total factor productivity is persistent over time, distinguishing between the cases in which ω_{it} follows linear and non-linear dynamic processes. In both cases the dynamic process for ω_{it} has to be correctly specified by the researcher.

Linear TFP process

The moment conditions discussed above extend straightforwardly to cases in which ω_{it} follows a low order ARMA process. Suppose, for example, that ω_{it} follows an AR(1) process

$$\omega_{it} = \rho\omega_{i,t-1} + \upsilon_{it} \tag{11}$$

with $|\rho| < 1$, in which the productivity innovations υ_{it} are independent across firms and serially uncorrelated. Substituting for ω_{it} and $\omega_{i,t-1}$ in (11) from (9) results in a quasi-differenced representation of the quantity production function in which the error term is now $u_{it} := \upsilon_{it} + \varepsilon_{it} - \rho\varepsilon_{i,t-1}$:

$$(y_{it} - \beta_K k_{it} - \beta_L l_{it} - \beta_M m_{it} - \varepsilon_{it}) = \rho(y_{i,t-1} - \beta_K k_{i,t-1} - \beta_L l_{i,t-1} - \beta_M m_{i,t-1} - \varepsilon_{i,t-1}) + \upsilon_{it}$$

$$\Leftrightarrow (y_{it} - \rho y_{i,t-1}) = \beta_K(k_{it} - \rho k_{i,t-1}) + \beta_L(l_{it} - \rho l_{i,t-1}) + \beta_M(m_{it} - \rho m_{i,t-1}) + (\upsilon_{it} + \varepsilon_{it} - \rho\varepsilon_{i,t-1}).$$

Here we still have moment conditions of the form $\mathrm{E}[(k_{is}, l_{is})u_{it}] = 0$ for $s \leqslant t$. If the researcher has data on the input price, and the input price is uncorrelated with ω_{it}, we have additional moment conditions $\mathrm{E}[p^M_{is}u_{it}] = 0$ for $s \leqslant t$. If the researcher does not have data on the input price, or if the variation across firms in these input prices is correlated with ω_{it}, but we have persistent variation across firms in either p^M_{it} or ξ_{it}, we have additional (informative) moment conditions $\mathrm{E}[m_{is}u_{it}] = 0$ for $s \leqslant t-1$. If the measurement error ε_{it} is serially uncorrelated, we also have additional moment conditions $\mathrm{E}[y_{is}u_{it}] = 0$ for $s \leqslant t-2$.[19] These moment conditions can be used to estimate the parameter vector $(\beta_K, \beta_L, \beta_M, \rho)$ consistently in the quasi-differenced quantity production function, following the approach suggested by Blundell and Bond (2000).

Non-linear TFP process

Similar moment conditions could be used to estimate the output elasticity parameters consistently with non-linear processes for ω_{it}, if gross output is measured without error and ω_{it} is the only component of the error term in the quantity production function (9). Otherwise, if we replace the linear AR(1) process (11) by the first-order Markov process

$$\omega_{it} = g(\omega_{i,t-1}) + \upsilon_{it}\,, \tag{12}$$

the presence of the unobserved $\varepsilon_{i,t-1}$ inside the non-linear function $g(\omega_{i,t-1})$, when we substitute for $\omega_{i,t-1}$ using (9), will invalidate moment conditions of the kind considered in the previous sub-section.

With a non-linear process for ω_{it} and measurement error in output, Flynn et al. (2019) have shown that when firms have market power, even with a quantity measure of output and all inputs, gross output production functions with a flexible input are not identified if the decision rule for the flexible input has the form $m_{it} = m^*_t(k_{it}, l_{it}, \omega_{it})$. Comparison to the decision rule for m_{it} in (8) indicates that this assumption rules out variation across firms in both the input price (p^M_{it}) and the demand shifter (ξ_{it}).[20] These are the sources of variation that we relied on in the previous sub-section, for identification of the output elasticity for the flexible input (β_M) in specifications with linear processes for ω_{it}. Our contribution in this sub-section is to consider whether variation across firms in either p^M_{it} or ξ_{it} would allow this key output elasticity parameter to be estimated consistently, in specifications with a non-linear process for ω_{it} and measurement error in output.

[17] This can also be seen from the decision rule for m_{it} given in (8). We assume here that the researcher does not observe the demand shifter. If the researcher observes ξ_{it}, and ξ_{it} varies across firms in a way which is uncorrelated with ω_{it}, then ξ_{it} could be used as an instrument for m_{it} in (9), and we would not require the variation across firms in ξ_{it} to be persistent. The same would apply if the researcher observes an informative proxy for ξ_{it} that is uncorrelated with ω_{it}.

[18] See Bond and Söderbom (2005), Ackerberg et al. (2015) and Gandhi et al. (2020).

[19] This restriction follows naturally if the ε_{it} component of the error term in (9) is interpreted as a shock to productivity that is not known by the firm when making input decisions in period t.

[20] The dependence of $m^*_t(k_{it}, l_{it}, \omega_{it})$ on the time period t allows for common variation over time in both p^M_{it} and ξ_{it}.

We still have moment conditions of the form $\mathrm{E}[(k_{is}, l_{is})\upsilon_{it}] = 0$ for $s \leqslant t$ and, for example, $\mathrm{E}[m_{is}\upsilon_{it}] = 0$ for $s \leqslant t-1$. To exploit these moment conditions, we would first need to eliminate the measurement error component ε_{it} from the error term of the quantity production function (9), before we substitute for $\omega_{i,t-1}$ in the non-linear function $g(\omega_{i,t-1})$.

A two stage estimation procedure of this kind was proposed by Ackerberg et al. (2015) for the estimation of a value added production function for price-taking firms, and with no flexible inputs. Similar two stage estimators are commonly used in the empirical literature that uses the ratio estimator to study markups.[21] De Loecker and Warzynski (2012) proposed an estimator of this type which can be used when we observe firm-level prices for both output and the flexible input, and we have persistent variation across firms in the price of the flexible input, and no unobserved variation across firms in the demand shifter. However, there are problems in applying this approach to the estimation of a gross output production function when firms have market power and there is unobserved heterogeneity across firms in the demand shifter ξ_{it}.

The first stage of these two stage procedures relies on having a valid control function which expresses the unobserved ω_{it} in (9) as a function of observed variables only. This is obtained by expressing the firm's optimal choice of the flexible input m_{it} as a function of observed variables and the single unobserved component ω_{it}. We also require that this function is strictly monotonic in ω_{it}, so that it can be inverted to provide the control function. A (possibly non-parametric) regression of y_{it} on the observed inputs and any additional observed variables included in the control function then has the error term ε_{it}. The predicted values of y_{it} from the estimated first stage regression can then be used in place of the actual values of y_{it} when we substitute for ω_{it} and $\omega_{i,t-1}$ in the specified non-linear dynamic process (12).

The question here is whether we can find a valid control function of this form in settings where we also have informative instruments for m_{it} in the second stage of this procedure. We have the decision rule $m_{it} = m^*(k_{it}, l_{it}, \omega_{it}, \xi_{it}, p^M_{it})$ obtained in (8). First suppose that the researcher has data on p^M_{it} and all firms face the same demand curve ($\xi_{it} = \xi_t$ for all i). Time dummies (d_t) can then be used to control for the common demand shocks. The decision rule then depends on the scalar unobservable ω_{it}, and can be inverted to give the valid control function $\omega_{it} = h(k_{it}, l_{it}, m_{it}, p^M_{it}, d_t)$, which can be used in the first stage regression. If the variation in p^M_{it} is uncorrelated with ω_{it}, we can also use the observed input prices as instruments for m_{it} in the second stage specification; that is, we have valid and informative moment conditions of the form $\mathrm{E}[p^M_{is}\upsilon_{it}] = 0$ for $s \leqslant t$. If the variation in p^M_{it} is correlated with ω_{it} but persistent over time, we can instead use lagged intermediate inputs as instruments for m_{it} in the second stage specification; that is, we have valid and informative moment conditions of the form $\mathrm{E}[m_{is}\upsilon_{it}] = 0$ for $s \leqslant t-1$. Notice that with no heterogeneity across firms in the demand shifter, we require persistent variation across firms in the input price here; with firm-level data on the input price, this condition can be checked.

Now suppose that the researcher has data on p^M_{it} and there is variation across firms in the demand shifter which is not *perfectly* observed by the researcher (i.e. there is *some* unobserved heterogeneity across firms in ξ_{it}). In this case, we can no longer express m_{it} as a function of observed variables and the scalar unobservable ω_{it}. We could still invert the function $m_{it} = m^*(k_{it}, l_{it}, \omega_{it}, \xi_{it}, p^M_{it})$ to obtain $\omega_{it} = h(k_{it}, l_{it}, m_{it}, \xi_{it}, p^M_{it})$, but this does not provide a valid control function for ω_{it} if there is any unobserved variation across firms in the demand shifter ξ_{it}.

Similar issues arise if we consider using the first order condition (10) as the basis for obtaining a control function for ω_{it} in the first stage regression. In this case, we could still invert the function $m_{it} = m(k_{it}, l_{it}, \omega_{it}, z_{it}, p^M_{it})$ to obtain $\omega_{it} = h(k_{it}, l_{it}, m_{it}, z_{it}, p^M_{it})$, but with unobserved heterogeneity in ξ_{it}, the researcher would need to be able to control for variation in the log of marginal cost z_{it}, to obtain a valid control function.[22] Otherwise, with market power and unobserved heterogeneity in demand, we cannot allow for non-linearity in the dynamic process for total factor productivity using a two stage procedure of this type, even with firm-level data on the price of the flexible input.[23]

3.1.3. Data on revenue and output price indices

The previous section considered the case in which the researcher has data on both sales revenue and the *level* of the output price for individual firms. An intermediate possibility is that the researcher observes an output price *index* for individual firms, constructed from survey questions about yearly price changes, but does not observe firm-specific price levels in the base year.

If we use these firm-specific output price indices to deflate the value of output in current prices, we obtain

$$P_{i0}Q_{it} := (P_{it}Q_{it}) \times \left(\frac{P_{it}}{P_{i0}}\right)$$

where (P_{it}/P_{i0}) is the firm-specific price index, equal to one in the base period $t = 0$, and P_{i0} is the unobserved price of output for firm i in that period.

[21] See, for example, De Loecker and Warzynski (2012) and De Loecker et al. (2020).

[22] This has also been noted by Doraszelski and Jaumandreu (2019) in a more general setting than our example here. With no unobserved variation across firms in ξ_{it}, we have $z_{it} = z(k_{it}, l_{it}, \omega_{it}, p^M_{it}, d_t)$. Substituting for z_{it} in the first order condition and inverting the resulting function then gives the same control function $\omega_{it} = h(k_{it}, l_{it}, m_{it}, p^M_{it}, d_t)$ that we obtained from the decision rule.

[23] The situation is no better if the researcher does not have data on the price of the flexible input. To obtain a valid control function for ω_{it} in the first stage regression, we then require no unobserved heterogeneity across firms in ξ_{it} and no variation across firms in p^M_{it}. Observed variation in the demand shifter ξ_{it} would then be needed to provide informative instruments for m_{it} in the second stage specification, and this approach could not be used in a specification with two or more flexible inputs.

Deflating revenue in this way measures the true level of output Q_{it} up to the unknown multiplicative firm-specific constant P_{i0}, reflecting unobserved differences across firms in the price of output in the base year. In a logarithmic specification, this will introduce firm-specific intercepts. For example, for the Cobb-Douglas gross output production function considered in the previous section, we obtain from (9)

$$(p_{i0} + y_{it}) = p_{i0} + \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + (\omega_{it} + \varepsilon_{it}) \tag{13}$$

where again $y_{it} = q_{it} + \varepsilon_{it}$, and ε_{it} allows for transient measurement error. Persistent differences across firms in the level of the output price will be correlated with input choices, so in the panel data sense these firm-specific intercepts will need to be treated as 'fixed effects' (i.e. correlated with the explanatory variables) rather than 'random effects' (i.e. uncorrelated with the explanatory variables).[24]

In the case where the unobserved total factor productivity component of the error term ω_{it} follows a low order ARMA process, the 'dynamic panel data' estimator for production functions proposed by Blundell and Bond (2000) can accommodate unobserved firm-specific fixed effects of this form. This allows consistent estimation of the output elasticity parameters $(\beta_K, \beta_L, \beta_M)$ provided that ω_{it} follows a linear process and either: (i) we have data on p^M_{it}, and the input price is uncorrelated with ω_{it}; or (ii) there is persistent variation across firms in either p^M_{it} or ξ_{it}, such that lagged inputs provide valid and informative instruments for m_{it}. The key point here is that estimation will need to allow for fixed effects if the researcher does not have firm-level data on output price levels.

The two stage estimators which have been developed to allow for non-linear dynamics in ω_{it} cannot allow for unobserved firm-specific fixed effects in ω_{it}, at least in panel data settings with a small number of time periods. It may be possible to extend estimators of this type to allow for unobserved firm-specific fixed effects in the measurement error component of the error term, which is the relevant case here. This could be a useful subject for further research, in settings where we have data on firm-specific price indices but not firm-specific price levels, and are content with the assumption of no unobserved variation across firms in the demand shifter ξ_{it}.

3.2. *Estimation of the revenue elasticity for a flexible input*

In Section 3.1 we showed that the output elasticity for a flexible input is not identified from estimation of the revenue production function without strong parametric restrictions on the forms of both the gross output production function and the inverse demand curve. In this section, we briefly consider conditions under which the revenue elasticity for a flexible input can be estimated consistently.

A useful starting point is the case considered by Klette and Griliches (1996), with a Cobb-Douglas gross output production function (9) and a CES inverse demand curve

$$p_{it} = \delta_t - \eta^{-1} q_{it} + \zeta_{it} \tag{14}$$

in which we have decomposed the demand shifter ξ_{it} into common and idiosyncratic components, such that $\xi_{it} = \delta_t + \zeta_{it}$. Here $\eta > 1$ is the absolute value of the price elasticity of demand. The revenue production function in this case is

$$r^o_{it} = (p_{it} + y_{it}) = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + (p_{it} + \omega_{it} + \varepsilon_{it}) \tag{15}$$

where the log of observed revenue $r^o_{it} := r_{it} + \varepsilon_{it}$ differs from the log of true revenue r_{it} by the additive measurement error component ε_{it}.

Substituting for the unobserved output price p_{it} in the error term of (15) from the inverse demand curve (14), we obtain the log-linear equation

$$r^o_{it} = \delta_t + \left(\frac{\beta_K}{\mu}\right) k_{it} + \left(\frac{\beta_L}{\mu}\right) l_{it} + \left(\frac{\beta_M}{\mu}\right) m_{it} + \left[\left(\frac{1}{\mu}\right)\omega_{it} + \zeta_{it} + \varepsilon_{it}\right] \tag{16}$$

which relates observed revenue to the observed inputs. Here $\mu = \left(1 - \frac{1}{\eta}\right)^{-1} > 1$ is the markup, and the slope parameters are the *revenue* elasticities. The error term contains the idiosyncratic demand shock ζ_{it}, in addition to total factor productivity ω_{it} and the measurement error ε_{it}.

The revenue elasticity parameters in (16) can then be estimated consistently using the methods discussed in Section 3.1.2, subject to the limitations that we have noted. For example, if both ω_{it} and ζ_{it} are assumed to be serially uncorrelated, we have moment conditions $E[(k_{it}, l_{it})u_{it}] = 0$, where now $u_{it} := \left(\frac{1}{\mu}\right)\omega_{it} + \zeta_{it} + \varepsilon_{it}$. With persistent variation across firms in the input price p^M_{it}, the lagged input $m_{i,t-1}$ provides a valid and informative instrument for m_{it}, and we have the additional (informative) moment condition $E[m_{i,t-1}u_{it}] = 0$. This extends straightforwardly to cases in which ω_{it} follows a low order ARMA process, although not to cases in which ω_{it} follows a non-linear dynamic process (if we do indeed have both unobserved idiosyncratic demand shocks and measurement error).

[24] A similar issue arises if we use an expenditure measure of one or more of the inputs, deflated using a firm-specific input price index, and there is unobserved variation across firms in the level of the input price.

In cases where we can estimate these revenue elasticity parameters consistently, we could investigate heterogeneity in the markup parameter μ across (large) sub-samples of firms by including suitable interaction terms in (16), under the maintained assumption that the output elasticities are common to these sub-samples.[25]

This example also highlights potential problems with estimating the revenue elasticities consistently. Consistent estimation in the example considered above required the researcher to observe a quantity measure of the flexible input.[26] More generally, consistent estimation may be difficult if the sum $\left(\frac{1}{\mu}\right)\omega_{it} + \zeta_{it}$ does not follow a low order ARMA process. Consistent estimation may also be difficult if the markup parameter μ is not common within (large) sub-samples of firms. The moment conditions that are typically used to estimate production functions will not be valid if there is unmodeled heterogeneity in the slope parameters in (16).[27] Finally, consistent estimation of the revenue elasticities is likely to be more difficult if the gross output production function and inverse demand curve do not take the convenient log-linear forms implied by a Cobb-Douglas production technology and a CES demand schedule.

4. Conclusion

Our primary objective in this paper is to caution against drawing inferences from firm-level markup estimates based on the production approach, when firm-level output prices are not observed. Static profit maximization conditions imply that when a revenue elasticity is used in place of an output elasticity, the commonly-used ratio estimator contains no useful information about markups. Static profit maximization also implies that the required output elasticity for a flexible input is not identified from estimation of a revenue production function, without placing strong parametric restrictions on the functional forms of both the production function and the demand schedule. We discuss additional problems with the ratio estimator of markups when the flexible input is used by firms not only to produce output, but also to influence demand; and we show that even with separate data on output prices and quantities, it is still challenging to estimate the output elasticity for a flexible input consistently, if there are non-linear productivity dynamics and firms face heterogeneous demand schedules, with unobserved heterogeneity across firms in a demand shifter.

These difficulties notwithstanding, the clear implication of our main results is that firm-level data on output prices are required to obtain credible estimates of markups using the ratio estimator. With revenue data alone, we are not aware of any procedures that would allow the level of markups to be recovered, without imposing additional structure on the demand side of the market. If a researcher is reluctant to place structure on demand, an alternative is to focus instead on the difference in mean markups between groups of firms for which one is comfortable assuming that the production function parameters are the same across the groups. In Appendix B.2 we show that this difference can be estimated consistently without knowledge of the output elasticity, using a regression specification for the cost share in revenue for a flexible input. A leading example would be the comparison of mean markups across exporters and non-exporters, considered in De Loecker and Warzynski (2012), provided one is willing to assume that production function elasticities do not vary systematically with export status. However, this approach is not well suited to studying trends in markups, since the maintained assumption that the output elasticity is stable over time cannot be verified without a way of estimating the output elasticity consistently for different sub-periods of the sample.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jmoneco.2021.05.004.

References

Ackerberg, D.A., Caves, K., Frazer, G., 2015. Identification properties of recent production function estimators. Econometrica 83 (6), 2411–2451.
Baqaee, D.R., Farhi, E., 2020. Productivity and misallocation in general equilibrium. Q. J. Econ. 135 (1), 105–163.
Basu, S., 2019. Are price-cost markups rising in the united states? A discussion of the evidence. J. Econ. Perspect. 33 (3), 3–22.
Blundell, R., Bond, S., 2000. GMM estimation with persistent panel data: An application to production functions. Econom. Rev. 19 (3), 321–340.
Bond, S., Söderbom, M., 2005. Adjustment Costs and the Identification of Cobb-Douglas Production Functions. Working Paper No. 05/04. Institute for Fiscal Studies,.
Burstein, A., Carvalho, V.M., Grassi, B., 2020. Bottom-Up Markup Fluctuations. Working Paper No. 27958. National Bureau of Economic Research,.
De Loecker, J., 2011. Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. Econometrica 79 (5), 1407–1451.

[25] For example, we could investigate if the revenue elasticity parameters take different values for exporting and non-exporting firms, as in De Loecker and Warzynski (2012).

[26] If the researcher only has data on expenditure on the flexible input, the assumption that the price of the flexible input does not vary across firms then implies that the lagged input is not an informative instrument for the current input, given the levels of the predetermined inputs k_{it} and l_{it}, under the maintained assumptions that ω_{it} and ζ_{it} are both serially uncorrelated; see (10).

[27] In the model $y_{it} = \beta x_{it} + u_{it}$ with $\mathrm{E}(u_{it}) = 0$ and $\mathrm{E}(x_{it}u_{it}) \neq 0$, we can obtain consistent estimators of β if $\mathrm{E}(x_{i,t-1}u_{it}) = 0$ and $x_{i,t-1}$ is also an informative instrument for x_{it}. With heterogeneity across firms in the slope parameter, we have $y_{it} = \beta_i x_{it} + u_{it} = \beta x_{it} + u_{it} + (\beta_i - \beta)x_{it} = \beta x_{it} + e_{it}$, with $e_{it} := u_{it} + (\beta_i - \beta)x_{it}$. If the explanatory variable is serially correlated, we then have $\mathrm{E}(x_{i,t-1}e_{it}) \neq 0$, and standard estimators do not estimate β consistently. With time-invariant heterogeneity of this form, the β_i coefficients (and hence β) could be estimated consistently if panel data is available for a large number of time periods. See Pesaran and Smith (1995) for further discussion.

De Loecker, J., Eeckhout, J., Unger, G., 2020. The rise of market power and the macroeconomic implications. Q. J. Econ. 135 (2), 561–644.
De Loecker, J., Goldberg, P.K., 2014. Firm performance in a global market. Annu. Rev. Econ. 6 (1), 201–227.
De Loecker, J., Warzynski, F., 2012. Markups and firm-level export status. Am. Econ. Rev. 102 (6), 2437–2471.
Doraszelski, U., Jaumandreu, J., 2019. Using Cost Minimization to Estimate Markups. Discussion Paper No. DP14114. CEPR,.
Flynn, Z., Traina, J., Gandhi, A., 2019. Measuring Markups with Production Data. Working Paper. https://ssrn.com/abstract=3358472,.
Gandhi, A., Navarro, S., Rivers, D.A., 2020. On the identification of gross output production functions. J. Polit. Economy 128 (8), 2973–3016.
Hall, R.E., 1986. Market structure and macroeconomic fluctuations. Brookings Pap. Econ. Act. 17 (2), 285–338.
Hall, R.E., 1988. The relation between price and marginal cost in US industry. J. Polit. Economy 96 (5), 921–947.
Klette, T.J., Griliches, Z., 1996. The inconsistency of common scale estimators when output prices are unobserved and endogenous. J. Appl. Econom. 11 (4), 343–361.
Mrázová, M., Neary, J.P., Parenti, M., 2018. Sales and Markup Dispersion: Theory and Empirics. Working Paper No. 7433. CESifo,.
Pesaran, M.H., Smith, R., 1995. Estimating long-run relationships from dynamic heterogeneous panels. J. Econom. 68 (1), 79–113.
Syverson, C., 2011. What determines productivity? J. Econ. Lit. 49 (2), 326–365.
Syverson, C., 2019. Macroeconomics and market power: context, implications, and open questions. J. Econ. Perspect. 33 (3), 23–43.
Traina, J., 2018. Is Aggregate Market Power Increasing? Production Trends using Financial Statements. Working Paper No. 17. Stigler Center,.