NBER WORKING PAPER SERIES

SOME UNPLEASANT MARKUP ARITHMETIC: PRODUCTION FUNCTION ELASTICITIES AND THEIR ESTIMATION FROM PRODUCTION DATA

Steve Bond
Arshia Hashemi
Greg Kaplan
Piotr Zoch

Working Paper 27002
http://www.nber.org/papers/w27002

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2020

We thank Tugce Turk for excellent research assistance. We thank Susanto Basu, Chad Syverson and Ali Hortacsu for helpful suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w27002.ack

Some Unpleasant Markup Arithmetic: Production Function Elasticities and their Estimation from Production Data

Steve Bond, Arshia Hashemi, Greg Kaplan, and Piotr Zoch

NBER Working Paper No. 27002
April 2020
JEL No. D2,D4,L1,L4

ABSTRACT

The ratio estimator of a firm's markup is the ratio of the output elasticity of a variable input to that input's cost share in revenue. This note raises issues that concern identification and estimation of markups using the ratio estimator. Concerning identification: (i) if the revenue elasticity is used in place of the output elasticity, then the estimand underlying the ratio estimator does not contain any information about the markup; (ii) if any part of the input bundle is either used to influence demand, or is neither fully fixed nor fully flexible, then the estimand underlying the ratio estimator is not equal to the markup. Concerning estimation: (i) even with data on output quantities, it is challenging to obtain consistent estimates of output elasticities when firms have market power; (ii) without data on output quantities, as is typically the case, it is not possible to obtain consistent estimates of output elasticities when firms have market power and markups are heterogeneous. These issues cast doubt over whether anything useful can be learned about heterogeneity or trends in markups, from recent attempts to apply the ratio estimator in settings without output quantity data.

Steve Bond
Nuffield College
New Road,
Oxford
OX1 1NF
United Kingdom
steve.bond@nuffield.ox.ac.uk

Greg Kaplan
Department of Economics
University of Chicago
1126 E 59th St
Chicago, IL 60637
and NBER
gkaplan@uchicago.edu

Arshia Hashemi
Department of Economics
University of Chicago
1126 E 59th St
Chicago, IL 60637
arshiahashemi@uchicago.edu

Piotr Zoch
Department of Economics
University of Chicago
1126 E 59th St
Chicago, IL 60637
pzoch@uchicago.edu

1 Introduction

This paper is about the interpretation of estimates of firm-level markups based on the production function approach. Under this approach, the estimator of the markup is the ratio of the output elasticity of a variable input to that input's cost share in revenue. We refer to this estimator of the markup as the *ratio estimator*. The production function approach was pioneered by Hall (1988, 1986), in his estimates of industry-level markups. The ratio estimator builds on Hall's ideas and has recently been used to estimate firm-level markups by De Loecker and Warzynski (2012), De Loecker et al. (2020) and many others. The resulting estimates have received wide-spread attention and many potential issues in the interpretation of these estimates have already been discussed (see Traina (2018), Basu (2019), Syverson (2019), De Loecker and Eeckhout (2018)). The issues that we raise in this note appear to have been largely overlooked by the literature.

The issues we discuss are most relevant when data on output quantities are not available, as in the firm-level studies cited above. When output quantities are not available, it is common to proxy output with sales or value added, deflated with common industry-level price deflators. This approach effectively uses the *revenue* elasticity in place of the *output* elasticity in the numerator of the ratio estimator. Klette and Griliches (1996) show that when firm-level prices are correlated with input choices, the estimate of the output elasticity that is obtained in this way is biased downward. We show that for identifying and estimating markups, this problem is much more severe than just generating a downward bias in the ratio estimator. At least under monopolistic competition, whenever the true markup is different from one (i.e. price differs from marginal cost), the estimand underlying this version of the ratio estimator is not actually a function of the markup. Hence a ratio estimator that uses the revenue elasticity in the numerator contains no useful information about the markup at all, and the estimand underlying this estimator is identically equal to one.

In this paper, we pursue the implications of this observation and what they imply for identification and estimation of markups using the ratio estimator.

The first part of the paper concerns issues related to identification. In Section 2.1, we consider a best-case scenario in which all the assumptions needed for the ratio estimator to recover the markup from the output elasticity are satisfied, and in which the revenue and output elasticities are known. The main takeaway from this section is that it is essential to use the output elasticity, rather than the revenue elasticity, in the numerator of the ratio estimator in order to learn about markups. Even in this best-case scenario, replacing the output elasticity with the revenue elasticity removes all information about the markup from the ratio estimator. In Section 2.2, we raise two additional challenges for learning about markups that arise even if the output elasticity were known. 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 markups. 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 tasks (see Syverson (2011) for a related discussion in the context of productivity estimation).

The second part of the paper concerns issues related to estimation of the output elasticity that is needed in order for the ratio estimator to recover the markup. In Section 3.2 we show that even if data on output quantities and input quantities are available, it is still challenging to obtain consistent estimates of output elasticities for flexible inputs. In particular, the variants of the Ackerberg et al. (2015) estimator that are typically used in the existing literature are not valid when firms have market power and residual demand schedules are heterogeneous, even with data on quantities. In Section 3.3 we then show that in the more usual setting, where only the value of output and expenditure on inputs are observed, it is not possible to obtain consistent estimates of the required *output* elasticity, when firms have market power and there is any interesting heterogeneity in markups. Finally, when data are available for many firms, but only for a small number of time periods, we show that it is not even possible to obtain consistent estimates of the *revenue* elasticity, if there is firm-level heterogeneity in markups.

Overall, the identification and estimation issues that we raise cast serious doubt over whether anything useful can be learned about trends or heterogeneity in markups from the ratio estimator, unless firm-level data on output quantities are observed.

2 Difficulties in Recovering Markups from Production Function Elasticities

In this section, we clarify conditions under which markups can be recovered from knowledge of production function elasticities and input cost shares in total revenue. We first emphasize that knowledge of the *output* elasticity with respect to a flexible input, as opposed to the *revenue* elasticity, is essential in this regard. We then mention additional implicit assumptions that are required to recover markups even if output elasticities are known. Throughout this section, we abstract from firm heterogeneity and stochastic shocks; we consider these features in Section 3 where we discuss challenges to estimating the elasticities that treated as known in this section.

2.1 Revenue elasticities versus output elasticities

Consider a firm that produces output Q using a production function with N inputs, X_i, $i = 1 \dots N$.

$$Q = F_Q\left(X_1, X_2, \dots\right)$$

The firm purchases inputs in perfectly competitive markets at prices W_i, which it takes as given.[1] The firm faces an inverse demand curve $P(Q)$; its total revenue is given by $R(Q) = P(Q)Q$. Note that the elasticity of revenue with respect to an input X_i is determined by both the elasticity of the inverse demand curve $\varepsilon_{P,Q} := \frac{\partial P}{\partial Q}\frac{Q}{P}$ and the output elasticity of the input $\varepsilon_{Q,X_i} := \frac{\partial Q}{\partial X_i}\frac{X_i}{Q}$ as

$$\varepsilon_{R,X_i} = \left(1 + \varepsilon_{P,Q}\right)\varepsilon_{Q,X_i} \tag{1}$$

The profit maximization problem of the firm can be expressed as

$$\Pi = \max_{Q} P(Q)Q - C(Q), \tag{2}$$

where $C(Q)$ is the firm's cost function, defined by

$$\begin{aligned} C(Q) := &\min_{X_i} \sum_i X_i W_i \\ &\text{subject to} \\ Q \leq &F_Q\left(X_1, X_2, \dots\right) \end{aligned} \tag{3}$$

Attaching a Lagrange multiplier $\lambda \geq 0$ to the constraint in the cost minimization problem, yields the necessary conditions

$$\begin{aligned} W_i &= \lambda \frac{\partial}{\partial X_i} F_Q\left(X_i\right) \ \forall\ i \\ \frac{W_i X_i}{PQ} &= \frac{\lambda}{P}\varepsilon_{Q,X_i} \end{aligned}$$

Using s_{R,X_i} to denote the share of input i's cost in revenue and applying the envelope condition yields the familiar relationship between the price to marginal cost ratio, the

[1] For simplicity, we treat all inputs X_i as fully flexible inputs but this is not essential to the points we make in this section, since if a subset of inputs were fully fixed, we could work with the conditional cost function. In Appendix A, we show that if a subset of inputs is partially fixed and incurs adjustment costs that depend on the input choice, this would also not affect the non-identification result with revenue elasticities, and would introduce a bias even in the case where output elasticities were observed.

output elasticity, and the input cost share in revenue

$$s_{R,X_i} = \frac{C'(Q)}{P}\varepsilon_{Q,X_i} \tag{4}$$

The first-order condition for the profit maximization problem (2) implies

$$\frac{C'(Q)}{P} = 1 + \varepsilon_{P,Q} \tag{5}$$

so that the markup of price over marginal cost is given by $\mu := \frac{P}{C'(Q)} = (1 + \varepsilon_{P,Q})^{-1}$.

The production function approach to estimating markups is to use the ratio of the output elasticity of a variable input ε_{Q,X_i} to that input's cost share in revenue s_{R,X_i}. We will denote the estimand underlying the ratio estimator by $\hat{\mu}_Q := \frac{\varepsilon_{Q,X_i}}{s_{R,X_i}}$. Re-arranging (4) shows that

$$\hat{\mu}_Q = \mu$$

and so the ratio estimator correctly recovers the markup of price over marginal cost.

What does the ratio estimator recover if one uses the revenue elasticity in place of the output elasticity? We denote this estimand by $\hat{\mu}_R := \frac{\varepsilon_{R,X_i}}{s_{R,X_i}}$. Combining (1), (4) and (5) shows that

$$\hat{\mu}_R = 1$$

So using the revenue elasticity in place of the output elasticity only recovers an estimate of the markup when the true markup is 1, i.e. when price is equal to marginal cost. 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 quantity. But the ability to affect price by changing quantity is the typical reason why a firm would price at a markup over marginal cost. Since the estimand is identically equal to 1 when the revenue elasticity is used in place of the output elasticity, the ratio of the revenue elasticity to the cost share in revenue does not contain *any* information about the actual markup of price over marginal cost.

This observation is closely related to Klette and Griliches (1996), who showed that using revenue in place of output to estimate an output elasticity produces a downward bias. In our simple example, this effect is readily seen from equation (1), together with the typical assumption that demand curves slope downward $\varepsilon_{P,Q} < 0$. Since the ratio estimator uses the output elasticity in the numerator, Klette and Griliches (1996) is often cited as a reason why using revenue elasticities to estimate markups 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 1, it is the wrong interpretation of the result. The bias in the estimator is the only part of the estimator that contains any

information about the markup, so that the biased estimator is not informative about the markup at all.

Unfortunately, output Q is rarely observed for individual firms. Instead, researchers typically only have access to measures of revenues or sales R. As we explain in Section 3, it is not possible to learn about the output elasticity ε_{Q,X_i} from data on revenue when firms have market power, using existing methods (and it is challenging even with data on output Q). It follows that with only data on revenues, nothing at all can be learned from the ratio estimator to learn about the level of markups.

Finally, it is useful to bear in mind that if it were somehow possible to learn the output elasticity with only knowledge of the revenue elasticity, then it would not be necessary to use the ratio estimator. One could simply estimate both the output elasticity and the revenue elasticity and note from equations (1) and (5) that the ratio of the 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 can therefore not 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 work does not find estimates from this approach that are centered around one. In the following sub-section, we mention two additional sources of bias in the ratio estimator that are likely to be reflected in these estimates. Then in Section 3.4 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

The previous section showed that when the revenue elasticity is used in the numerator of the ratio estimator, the resulting estimand is equal to unity, and contains no information about the markup. But when the output elasticity is used in the numerator of the ratio estimator, the resulting estimand correctly recovers the markup. 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 partly used to influence demand.

Input adjustment costs For the ratio estimator to recover the markup, it is crucial that the input X_i whose output elasticity and cost share are combined is perfectly flexible. Alternatively, as explained in Basu (2019), X_i can 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_i that is partially adjustable, or using a bundle that contains partially adjustable inputs, then 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 i is associated with a baseline quantity $\overline{X_i}$ and that the firm incurs adjustment costs when it chooses a quantity of input $X_i \neq \overline{X_i}$. The baseline quantity $\overline{X_i}$ 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 $\kappa_i(X_i)$, which satisfies $\kappa_i\left(\overline{X_i}\right) = \kappa_i'\left(\overline{X_i}\right) = 0$. In Appendix A we show that the ratio estimator using the revenue elasticity recovers

$$\hat{\mu}_R = \frac{\varepsilon_{R,X_i}}{s_{R,X_i}} = 1 + \frac{\kappa_i'(X_i)}{X_i},$$

and the ratio estimator using the output elasticity recovers[2]

$$\hat{\mu}_Q = \frac{\varepsilon_{Q,X_i}}{s_{R,X_i}} = \mu\left[1 + \frac{\kappa_i'(X_i)}{X_i}\right]$$

Thus, even if the output elasticity to 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.

Inputs that influence demand The framework in the previous section assumed that the inputs X_i are all used to produce output rather than to influence demand. Assume instead that the firm's revenue is given by

$$R = P(Q, D)\, Q$$

[2]These formulas assume that observed input costs are $W_i X_i$ rather than $W_i X_i + W_i \kappa_i(X_i)$. If observed input costs also include the adjustment costs then we would obtain $\hat{\mu}_Q = \mu\left(\frac{X_i + \kappa_i'(X_i)}{X_i + \kappa_i(X_i)}\right)$ which also does not recover the true markup.

where D is a demand shifter that the firm can influence through the use of inputs according to the function

$$D = F_D\left(X_{1D}, X_{2D}, \dots\right),$$

where we have denoted the amount of input i used in production as X_{iQ} and the amount used in influencing demand as X_{iD}. We assume that we can observe only the total quantity of input i used by the firm $X_i = X_{iD} + X_{iQ}$. In Appendix B we show that the estimand underlying the ratio estimator becomes

$$\hat{\mu}_Q = \mu \frac{\varepsilon_{X_{iQ},X_i}}{1 + \frac{X_{iD}}{X_{iQ}}},$$

where ε_{X_{iQ},X_i} describes how an additional unit of X_i is allocated between X_{iD} and X_{iQ}. So if the variable input is only used for production and not to influence demand ($\varepsilon_{X_{iQ},X_i} = 1$, $X_{iD} = 0$) then the ratio estimator recovers the markup. But if some of the input is used to influence demand, and this component is not separated out, then the ratio estimator will be biased downward. If the firm uses a constant fraction of the input X_i for production, then $\varepsilon_{X_{iQ},X_i} = 1$ and the ratio estimator is biased downward. For example, if, over time, the input X_i 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 first show that even if data on output *quantities* and input *quantities* are available, it is still challenging to obtain consistent estimates of output elasticities for flexible inputs. In particular, the variants of the Ackerberg et al. (2015) estimator that are typically used in the existing literature are not valid when firms have market power and residual demand schedules are heterogeneous, even with data on quantities. We then show that in the more usual setting where only the *value* of output and *expenditure* on inputs is observed, it is not possible to obtain consistent estimates of the required output elasticity when firms have market power and heterogeneous markups. Finally, when data is available for many firms

but only a small number of time periods, we show that it is not even possible to obtain consistent estimates of the *revenue* elasticity, if there is firm-level heterogeneity in markups.

3.1 Setup

For ease of exposition, throughout this section we focus on a Cobb-Douglas production function for gross output (Q) with three inputs: capital (K), labor (L), and intermediate inputs (M). However, the issues we highlight apply for any continuously differentiable gross output production function (see Appendix C for details). For firm i in period t, the production function can be written in log-linear form as

$$y_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + \omega_{it} + \varepsilon_{it}, \tag{6}$$

where natural logarithms are denoted in lower case (e.g. $m_{it} = \ln M_{it}$). Here y_{it} is the log of observed output and ε_{it} is mean zero measurement error, so that $y_{it} = q_{it} + \varepsilon_{it}$, and ω_{it} is the log of total factor productivity, which is observed by the firm but not by the researcher. We choose units such that the mean of ω_{it} is zero.

We assume that the quantity of intermediate inputs M_{it} is chosen optimally in period t, after the firm has observed ω_{it}, and that intermediate inputs do not incur adjustment costs of any kind. The quantities of capital K_{it} and labor L_{it} used in period t are assumed to be chosen optimally in period $t-1$, after the firm has observed $\omega_{i,t-1}$ but before the firm has observed ω_{it}. These predetermined inputs may also be subject to costs of adjustment; if so, these adjustment costs take the form of payments to third parties, not foregone production, and do not depend on M_{it}, nor on M_{is} in any other time period. We assume that the measurement error ε_{it} is uncorrelated with the observed inputs (k_{is}, l_{is}, m_{is}) for any s, t, and is independent across firms.[3] The common slope parameters $(\beta_K, \beta_L, \beta_M)$ are the output elasticities. Our parameter of interest is the output elasticity for the flexible input m_{it}.

We assume that the firm is a price-taker in all input markets, but has market power in the product market. Thus, the firm chooses its input demand for M_{it} and output level Q_{it} in each period to maximize the difference between its revenue $P_{it}Q_{it}$ and the cost of intermediate inputs $P_{it}^M M_{it}$, taking as given the costs of employing capital and labor, and any associated adjustment costs. In Appendix C, we show that the optimal choice of

[3] An alternative interpretation of $(\omega_{it}, \varepsilon_{it})$ 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 a serially uncorrelated productivity shock that is not known by the firm when making input decisions in period t.

intermediate inputs satisfies the first order condition

$$m_{it} = \frac{\ln \beta_M}{1-\beta_M} + \left(\frac{\beta_K}{1-\beta_M}\right) k_{it} + \left(\frac{\beta_L}{1-\beta_M}\right) l_{it} + \left(\frac{1}{1-\beta_M}\right) \left(p_{it} - \ln \mu_{it} - p_{it}^M + \omega_{it}\right), \quad (7)$$

where μ_{it} is the markup of price over marginal cost as in Section 2. The only restriction that we place on the residual demand curve is that the output price P_{it} is a weakly decreasing function of gross output Q_{it}.

3.2 Estimation of the output elasticity with output quantity data

We start with the special case in which productivity ω_{it} is independent across firms and serially uncorrelated. We discuss extensions to persistent processes for ω_{it} separately for each of the estimation strategies we consider below.

In the context of price-taking firms ($\mu_{it} = 1$), there are two types of moment conditions that have been used to obtain GMM estimators of the parameter vector $\beta = (\beta_K, \beta_L, \beta_M)$:

1. Moment conditions of the form $\mathrm{E}[(k_{it}, l_{it}, m_{i,t-1}) v_{it}] = 0$, where $v_{it} = \omega_{it} + \varepsilon_{it}$, underpin the GMM estimator suggested by Blundell and Bond (2000). We refer to this as the BB approach.

2. Moment conditions of the form $\mathrm{E}[(k_{it}, l_{it}, m_{i,t-1}) \omega_{it}] = 0$ underpin the GMM estimator suggested by Ackerberg et al. (2015). We refer to this as the ACF approach. Exploiting these ACF moment conditions requires a way to eliminate the measurement error component ε_{it} from the error term in the observed production function. This is the role of the first stage regression in the ACF approach, which requires a valid proxy for unobserved ω_{it}.

Since the firm's optimal choice of intermediate inputs m_{it} depends on its productivity ω_{it} (see equation (7)), under both approaches we require a valid and informative instrument for m_{it}. Both the predetermined inputs (k_{it}, l_{it}) and the lagged value of the flexible input $m_{i,t-1}$ are chosen in period $t-1$ and hence are uncorrelated with ω_{it} (which is assumed here to be serially uncorrelated). Thus, under our stated assumptions, both types of moment conditions are valid.

Identification using either the BB or ACF moments then requires that $m_{i,t-1}$ is an informative instrument for m_{it} in the production function (6), a requirement that we now discuss.

BB approach: perfect competition For price-taking firms ($\ln \mu_{it} = 0$), with the output price common to all firms ($p_{it} = p_t$ for all i) and with serially uncorrelated ω_{it}, we can see from (7) that in order for $m_{i,t-1}$ to be an informative instrument for m_{it} at given levels of k_{it} and l_{it}, we require persistent variation across firms in the price of the flexible input p^M_{it}.[4] Identification of the output elasticity for a perfectly flexible input fails if there is no variation in the real price of that input, and is likely to be weak if the only source of correlation between m_{it} and the instrument $m_{i,t-1}$ comes through persistence in a common real input price ($p^M_{it} - p_{it} = p^M_t - p_t$ for all i), unless data are available for many time periods.[5] However, if there is indeed persistent variation across firms in the input price p^M_{it}, we can obtain a consistent estimate of β_M using the BB moment conditions, provided we have quantity data on both inputs and output.

Blundell and Bond (2000) extend this approach to allow for low-order ARMA processes for ω_{it}, and for a time-invariant firm fixed-effect component of ω_{it}.[6] The BB approach also extends straightforwardly to more general functional forms for the production function with Hicks-neutral productivity.[7] The limitation of the BB approach is that it does not extend to non-linear dynamic processes for ω_{it}, due to the presence of the measurement error component ε_{it} in the composite error term v_{it}.

BB approach: market power In the more relevant case in which firms have market power, then with a single perfectly flexible input, it is clear from (7) that we no longer require persistent variation across firms in the input price p^M_{it} in order for $m_{i,t-1}$ to be an informative instrument for m_{it}, as long as there is persistent firm-level variation in output prices p_{it} and/or markups μ_{it}. Heterogeneous demand schedules with idiosyncratic demand shifters typically ensure firm-level variation in both output prices and markups.[8] The case of a single flexible input is particularly favorable here. For example, if we had two perfectly

[4] Our discussion here, and except where noted below, follows the recent literature on the estimation of firm-level production functions in assuming that firm-specific data on input and output *prices* are not available. If data on p^M_{it} are available, and the variation in these input prices is uncorrelated with v_{it}, these prices could also be used as instruments for m_{it}.

[5] See Bond and Söderbom (2005) and Gandhi et al. (2020), respectively.

[6] If the measurement error ε_{it} is serially uncorrelated or follows a low-order MA process, suitably lagged values of observed output $y_{i,t-k}$ for some $k > 0$ can be used as additional instruments. This is particularly important in the case of persistent ω_{it} processes for the identification of additional persistence parameters.

[7] See Doraszelski and Jaumandreu (2019) for a recent extension to a specification with labor-augmenting productivity.

[8] One exception is the CES demand schedule, given by $P_{it} = P_t \left(\frac{Q_{it}}{Q_t}\right)^{-\frac{1}{\sigma_t}} \exp(\xi_{it})$, where ξ_{it} is the idiosyncratic demand shifter, (P_t, Q_t) are the aggregate price and quantity indices, and $\sigma_t > 1$ is the (possibly time-varying) constant elasticity of substitution. This demand schedule implies common variation in markups $\mu_{it} = \sigma_t/(\sigma_t - 1)$, but still gives rises to firm-level variation in output prices, which is sufficient for $m_{i,t-1}$ to be an informative instrument here, provided the demand variation is persistent.

flexible inputs, then variation arising from heterogeneity in p_{it} or μ_{it} would be common to both inputs, and persistent variation across firms in at least one of the input prices would then be required, as in the case without market power discussed above.

This discussion clarifies that with quantity data on inputs and output, it may be possible to consistently estimate the output elasticity for a perfectly flexible input using the BB moment conditions, provided that unobserved productivity ω_{it} follows a linear dynamic process. However, this is not the approach that has been taken in the empirical literature that has applied the production approach to estimate markups. Instead, most of these applications have followed De Loecker and Warzynski (2012) in using variants of the ACF moment conditions.

ACF approach: perfect competition The ACF estimator was developed to estimate the elasticity of *value added* with respect to *predetermined* inputs for *price-taking* firms. The advantage of the ACF estimator in this context is that it can allow for non-linear dynamic processes for unobserved ω_{it}. But here our interest is in estimating the elasticity of *gross output* with respect to a *flexible* input for firms with *market power*. We will explain why the ACF approach cannot be used to obtain consistent estimates of the output elasticity for a perfectly flexible input, even if data on output and input quantities are available, when there is non-linearity in the productivity process, and heterogeneity in firms' residual demand schedules that results in unobserved variation in output prices and/or heterogeneity in markups. It follows that the ACF approach cannot be used to estimate output elasticities for the purpose of measuring heterogeneity in markups if non-linearity in the productivity process is an important feature of the data.

First, recall that in order to allow for non-linearity in the dynamic process for ω_{it}, the first stage of the ACF estimator purges (asymptotically) observed output y_{it} of measurement error ε_{it}. This first stage requires a valid proxy for unobserved productivity ω_{it}. To obtain a proxy, ACF assume that the optimal choice of intermediate inputs, $m_{it} = m_t(k_{it}, l_{it}, \omega_{it})$, is a strictly monotonic function of the scalar ω_{it}, which can be inverted to express $\omega_{it} = h_t(k_{it}, l_{it}, m_{it})$. For this scalar monotonicity condition to hold in our example, note from equation (7) that we would require: (i) that the firm is a price-taker in the product market ($\ln \mu_{it} = 0$) in addition to input markets; and (ii) that both output and input prices are common to all firms, i.e. $p_{it}^M - p_{it} = p_t^M - p_t$ for all i. If these conditions hold, then a linear regression of y_{it} on the observed (k_{it}, l_{it}, m_{it}) and time dummies would suffice to decompose observed y_{it} into consistent estimates of its q_{it} and ε_{it} components in the Cobb-Douglas case; more generally, a non-parametric regression of y_{it} on (k_{it}, l_{it}, m_{it}) for each period could be used.

This decomposition allows the elasticity of value added with respect to predetermined inputs for price-taking firms to be estimated consistently, allowing for non-linearity in the dynamic process for ω_{it}, if we have data on input and output quantities. But it does not follow that this approach can be used to obtain consistent estimates of the output elasticity for a flexible input, even in this setting. For convenience, we are assuming here that ω_{it} is serially uncorrelated. The natural application of the ACF estimator to the gross output production function then uses the moment conditions $\mathrm{E}[(k_{it}, l_{it}, m_{i,t-1})\omega_{it}] = 0$. However, the validity of the ACF proxy requires that the real input price is common to all firms, which means that, conditional on k_{it} and l_{it}, the only source of correlation between m_{it} and the instrument $m_{i,t-1}$ then comes through persistence in the common real input price $p_t^M - p_t$. Hence, identification is likely to be weak unless data are available for many time periods, and identification fails if the production function specification includes time-specific intercepts for any other reason.[9]

ACF approach: market power In the more relevant case in which firms have market power, then even the decomposition of observed output into actual output and measurement error fails, if residual demand schedules are heterogeneous. Even if we maintain that input prices are common across firms, with market power there is unobserved heterogeneity in the output price p_{it} and/or in the markup μ_{it}. From equation (7), we then have $m_{it} = m_t(k_{it}, l_{it}, z_{it}, \omega_{it})$, where $z_{it} := p_{it} - \ln \mu_{it}$ is the log of marginal cost. If all firms face the same residual demand schedule, we can express $z_{it} = z_t(k_{it}, l_{it}, \omega_{it})$, and hence express optimal m_{it} as a function of the scalar unobservable ω_{it}. But in the presence of any heterogeneity in demand schedules, z_{it} additionally depends on unobserved firm-level variation in the demand, violating the scalar unobservable condition required for the ACF proxy for ω_{it}. We can still invert $m_{it} = m_t(k_{it}, l_{it}, z_{it}, \omega_{it})$ to obtain $\omega_{it} = h_t(k_{it}, l_{it}, m_{it}, z_{it})$, but without data on marginal cost z_{it} we cannot use this expression to obtain a valid proxy for ω_{it}.

In Section 3.2 of their online appendix, De Loecker and Warzynski (2012) discuss an extension of the ACF proxy for ω_{it} to a setting where they assume both that there is persistent variation across firms in the price of the flexible input p_{it}^M, and that these firm-specific input prices are observed. In this case, the input demand condition becomes $m_{it} = m_t(k_{it}, l_{it}, z_{it}, p_{it}^M, \omega_{it})$ and we have $\omega_{it} = h_t(k_{it}, l_{it}, m_{it}, z_{it}, p_{it}^M)$. But even if we can observe input prices and include them in the first stage regression, this does not resolve the omission of the unobserved marginal cost term z_{it}. De Loecker and Warzynski (2012) also suggest including observed firm characteristics such as export status, which may be correlated with markups, as additional explanatory variables in the first stage regression specification. This

[9] See Bond and Söderbom (2005) and Gandhi et al. (2020) for further discussion. Ackerberg et al. (2015) recognized this difficulty in applying their estimator to gross output production functions (page 2428).

will only resolve the issue we highlight here if these additional firm characteristics are the *only* source of heterogeneity in the omitted z_{it} variable, at given levels of $(k_{it}, l_{it}, p_{it}^M, \omega_{it})$.

The same problem is also discussed in De Loecker et al. (2020), although they are also not clear about how z_{it} should then be measured. Implementation of this extended proxy for ω_{it} would require the researcher to have data on *marginal costs*, which itself is a function of output prices p_{it} and markups μ_{it}.[10] But one of the supposed attractions of the production approach to estimating markups is that it appears to not require measurement of marginal costs. There may be special cases in which we can relate the log of marginal cost to observed data on firms or establishments, but in general with demand heterogeneity we will not recover a valid proxy for unobserved ω_{it}. We thus conclude that it is not possible to allow for non-linearity in the productivity process using the ACF approach when firms have market power and residual demand schedules are heterogeneous. Hence neither of the standard estimation approaches allows the output elasticity for a perfectly flexible input to be estimated consistently in this setting, even with data on output and input quantities.

Estimation with only input expenditure data In practice, it is much more common to have data on sales revenue and expenditure on intermediate inputs, than it is to have data on output and input quantities.[11] When firms have market power and output prices are heterogeneous, the absence of data on output quantities presents a fundamental obstacle to the consistent estimation of output elasticities, which we review in the next sub-section. But the absence of data on input quantities poses less of a problem. Using data on the cost of intermediate inputs ($P_{it}^M M_{it}$), the gross output production function can be written as

$$y_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M e_{it} + \omega_{it} - \beta_M p_{it}^M + \varepsilon_{it},$$

where $e_{it} = p_{it}^M + m_{it}$ denotes the log of expenditure on intermediate inputs. The additional component of the error term $\beta_M p_{it}^M$ is problematic if input prices vary across firms, and especially problematic if this input price variation is persistent, in which case lagged values such as $e_{i,t-1}$ will not be valid instruments. This is usually addressed by assuming that the input price is common across firms ($p_{it}^M = p_t^M$ for all i), which may be a reasonable assumption if firms have no market power in input markets. In the Cobb-Douglas case,

[10] This has also been noted by Doraszelski and Jaumandreu (2019) in a more general setting than our example here.

[11] This also applies to the bundle of inputs represented by accounting data on the Cost of Goods Sold, which is assumed to be perfectly flexible in De Loecker et al. (2020). Note that we are abstracting here from any differences between sales and the value of production, and between purchases and the value of inputs used in production, due to changes in inventories. Further issues arise if these differences are material.

time-specific intercepts are then sufficient to control for this component of the error term.[12]

For price-taking firms ($\ln \mu_{it} = 0$) and a common output price p_t, identification then becomes weak because persistence in the common real input price ($p_t^M - p_t$) is the only source of correlation between the instrument $e_{i,t-1}$ and the endogenous variable e_{it}, conditional on k_{it} and l_{it}. For firms with market power, this is less problematic, at least in the case of a single flexible input, since persistent variation across firms in p_{it} or μ_{it} will again provide additional identifying information. The output elasticity parameters can then be estimated consistently using the BB moment conditions $\mathrm{E}[(k_{it}, l_{it}, e_{i,t-1})v_{it}] = 0$, and extensions of this approach to the case of low-order ARMA processes for ω_{it} are again possible.

3.3 Estimation of the output elasticity with revenue data

More fundamental problems arise when we attempt to use data on observed sales revenue $R_{it} \equiv P_{it}Y_{it}$ to estimate the output elasticity β_M. The revenue production function is

$$r_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M e_{it} + \omega_{it} + p_{it} - \beta_M p_{it}^M + \varepsilon_{it},$$

where $r_{it} = p_{it} + y_{it}$ denotes the log of observed revenue. As explained above, the input price component of the error term is not particularly problematic here if the input price is common, firms have market power, and use a single flexible input.

But the output price component of the error term p_{it} now presents a huge challenge to the consistent estimation of the output elasticity β_M, from direct estimation of this revenue production function, for firms with market power. Output prices will certainly vary across firms, even if input prices and the production technology are common. With heterogeneity in residual demand schedules, this variation in output prices reflects shocks to demand, as well as shocks to productivity. The output price also influences the optimal choice of intermediate inputs, and hence expenditure on those inputs. If this output price variation is serially uncorrelated, then with ω_{it} also serially uncorrelated, persistence in the common input price p_t^M is the only source of correlation between the instrument $e_{i,t-1}$ and the explanatory variable e_{it}, at given levels of k_{it} and l_{it} (see equation (7)). In this case, identification of the output elasticity parameters is again weak, unless data are observed for many time periods. Conversely, if the output price variation is persistent, this rules out using lagged input costs such as $e_{i,t-1}$ as instruments for e_{it} here. Consequently, it is not clear that the output elasticity β_M can be estimated consistently from the revenue

[12] Alternatively, input expenditure data in current prices can be deflated using a suitable price index for the common input price, and a single intercept is then sufficient. For more general functional forms, it may be necessary to express input expenditure data in constant (base year) prices.

production function, even in the simplest case where the productivity component of the error term ω_{it} is assumed to be serially uncorrelated.[13]

This problem is the omitted price bias that was highlighted by Klette and Griliches (1996), but which appears to have been either ignored or glossed over in most empirical applications of the production approach to estimating markups. No application that we are aware of can credibly claim to have obtained consistent estimates of the output elasticity, given the failure of the scalar unobservable condition that is needed to obtain a valid proxy for unobserved total factor productivity in the ACF approach, and the challenge of dealing with unobserved heterogeneity in the output price as an omitted variable.

3.4 Estimation of the revenue elasticity

A related question is whether it is possible to obtain consistent estimates of the *revenue* elasticity, using typical data on sales revenue and expenditure on intermediate inputs. One special case in which this may be possible is that studied by Klette and Griliches (1996), in which there is Cobb-Douglas technology, monopolistic competition, firms face (possibly idiosyncratic) constant elasticity of substitution demand schedules, and the markup is fixed across firms and over time ($\mu_{it} = \mu$). The revenue elasticity in this case is β_M/μ, consistent with our more general result in Section 2. If the output elasticity is common to all the firm-year observations in the sample, then the revenue elasticity is also a common parameter, which could be estimated consistently using panel data for a fixed number of time periods.[14] However, if there is unmodeled heterogeneity in the markup, the revenue elasticity is no longer a common parameter. All of the standard methods used to estimate (revenue) production functions rely on moment conditions of the form $\mathrm{E}(e_{i,t-1}v_{it}) = 0$ or $\mathrm{E}(e_{i,t-1}\omega_{it}) = 0$, which will not be valid if there is an additional error component due to unmodeled heterogeneity in the coefficient on e_{it} in the revenue production function.[15]

[13] One special case in which it may be possible to estimate the output elasticity β_M indirectly is that studied by Klette and Griliches (1996), with Cobb-Douglas technology and CES demands, that we discuss in sub-section 3.4 below. This case is of limited interest for learning about firm-level heterogeneity in markups.

[14] In this special case, we may be able to recover a consistent estimate of the output elasticity indirectly, from consistent estimates of the common revenue elasticity and demand elasticity (and hence markup) parameters. This estimate may be useful if we are interested in the technology, but contains no additional information if we are interested in the markup.

[15] 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 parameter, we have $y_{it} = \beta_i x_{it} + u_{it} = \beta x_{it} + u_{it} + (\beta_i - \beta)x_{it} = \beta x_{it} + \zeta_{it}$. If the explanatory variable is serially correlated, we then have $\mathrm{E}(x_{i,t-1}\zeta_{it}) \neq 0$, and standard estimators do not estimate β consistently. With time-invariant heterogeneity of this form, the β_i coefficients (and hence β) can be estimated consistently if panel data is available for a large number of time periods. See Pesaran and Smith (1995) for further discussion.

Thus with non-trivial heterogeneity across firms or time in markups, it is also challenging to estimate (mean) revenue elasticities consistently. No application that we are aware of in this literature has plausibly obtained consistent estimates of (mean) revenue elasticities. This may be an additional reason why the estimated markups are not centered around unity, as our analysis in Section 2 of the relationship between true revenue elasticities and cost shares in revenue for perfectly flexible inputs predicts.

4 Conclusion

Our objective with this note is to encourage others to exercise caution when drawing inferences from firm-level markup estimates based on the production function approach. We have shown that whenever a revenue elasticity is used in place of an output elasticity, at least under monopolistic competition, the commonly-used ratio estimator does not contain any useful information about markups. We are not aware of any procedures that would allow one to recover markups from revenue data alone, without imposing additional structure from the demand side of the market. We have also shown that violation of the widespread assumption that firms do not use inputs to influence their demand curves leads to an additional downward bias in the ratio estimator of markups. Since labor is used both to produce output and to influence demand, this suggests that labor should not be used as part of the input bundle when estimating markups. More generally, the assumption that *any* input bundle that contains a variable input can be used in the ratio estimator is too weak: it is also important that the input bundle does not contain any input that is used to influence demand.

Where does that leave us in terms of estimating firm-level markups? One possibility is to keep searching for reliable measures of changes in both price and quantity at the level at which one desires to estimate markups. This is the approach taken by Foster et al. (2008) for a small number of US manufacturing industries, and by Forlani et al. (2019) for Belgian manufacturing sectors in which units are well-defined. Another possibility is to estimate markups by estimating the demand elasticity directly, as in De Loecker (2011).

A third possibility is to give up on estimating the level of markups and focus on estimating the difference in mean markups across groups of firms for which one is comfortable with the assumption that they share the same production function parameters. This is the essence of the approach we outline in Appendix D. We show that for some questions about markups, one can work directly with the cost share in revenue of a variable input, and it is not necessary to use the ratio estimator. An example is the exercise in De Loecker and Warzynski (2012), in which they compare markups across exporters and non-exporters,

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.

References

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 2015, *83* (6), 2411–2451.

Basu, Susanto, "Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence," *Journal of Economic Perspectives*, 2019, *33* (3), 3–22.

Blundell, Richard W and Stephen R Bond, "GMM Estimation with Persistent Panel Data: An Application to Production Functions," *Econometric Reviews*, 2000, *19* (3), 321–340.

Bond, Stephen R and Måns Söderbom, "Adjustment Costs and the Identification of Cobb-Douglas Production Functions," Working Paper No. 05/04, Institute for Fiscal Studies, 2005.

De Loecker, Jan, "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity," *Econometrica*, 2011, *79* (5), 1407–1451.

_ **and Frederic Warzynski**, "Markups and Firm-Level Export Status," *American Economic Review*, 2012, *102* (6), 2437–71.

_ **and Jan Eeckhout**, "Some Thoughts on the Debate about (Aggregate) Markup Measurement," mimeo, 2018.

_ **, _ , and Gabriel Unger**, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, 2020, *forthcoming.*

Doraszelski, Ulrich and Jordi Jaumandreu, "Using Cost Minimization to Estimate Markups," Discussion Paper No. DP14114, CEPR, 2019.

Forlani, Emanuele, Ralf Martin, Giordiano Mion, and Mirabelle Muûls, "Unraveling Firms: Demand, Productivity and Markups Heterogeneity," Working Paper No. 5725, CESifo Group Munich, 2019.

Foster, Lucia, John Haltiwanger, and Chad Syverson, "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?," *American Economic Review*, 2008, *98* (1), 394–425.

Gandhi, Amit, Salvador Navarro, and David A Rivers, "On the Identification of Gross Output Production Functions," *Journal of Political Economy*, 2020, *forthcoming.*

Hall, Robert E, "Market Structure and Macroeconomic Fluctuations," *Brookings Papers on Economic Activity*, 1986, *17* (2), 285–338.

_ , "The Relation between Price and Marginal Cost in US Industry," *Journal of Political Economy*, 1988, *96* (5), 921–947.

Klette, Tor Jakob and Zvi Griliches, "The Inconsistency of Common Scale Estimators when Output Prices are Unobserved and Endogenous," *Journal of Applied Econometrics*, 1996, *11* (4), 343–361.

Pesaran, M Hashem and Ron Smith, "Estimating Long-Run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics*, 1995, *68* (1), 79–113.

Robinson, Peter M, "Root-N-Consistent Semiparametric Regression," *Econometrica*, 1988, *56* (4), 931–954.

Syverson, Chad, "What Determines Productivity?," *Journal of Economic Literature*, 2011, *49* (2), 326–65.

_ , "Macroeconomics and Market Power: Context, Implications, and Open Questions," *Journal of Economic Perspectives*, 2019, *33* (3), 23–43.

Traina, James, "Is Aggregate Market Power Increasing? Production Trends using Financial Statements," mimeo, University of Chicago, 2018.

Online Appendices

A Input adjustment costs

We consider the same firm problem from Section 2 but we now assume that each input i is associated with a baseline quantity $\overline{X_i}$ and that the firm incurs adjustment costs when it chooses a quantity of input $X_i \neq \overline{X_i}$. The baseline quantity $\overline{X_i}$ 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 $\kappa_i(X_i)$, which satisfies $\kappa_i\left(\overline{X_i}\right) = \kappa_i'\left(\overline{X_i}\right) = 0$.

The firm's cost function is now given by

$$\begin{aligned} C(Q) := &\min_{X_i} \sum_i X_i W_i + \sum_i \kappa_i(X_i) W_i \\ &\text{subject to} \\ Q \leq &F_Q(X_1, X_2, \dots) \end{aligned} \tag{8}$$

where we have normalized the adjustment cost functions by the input price W_i. Following the same steps as in the previous section, we obtain the FOC

$$\begin{aligned} W_i + W_i \kappa_i'(X_i) &= \lambda \frac{\partial}{\partial X_i} F_Q(X_i) \ \forall\ i \\ \frac{W_i X_i}{PQ}\left[1 + \frac{\kappa_i'(X_i)}{X_i}\right] &= \frac{\lambda}{P} \varepsilon_{Q,X_i} \end{aligned}$$

Using s_{R,X_i} to denote the share of input i's cost in revenue and using the envelope condition, this implies

$$s_{R,X_i}\left[1 + \frac{\kappa_i'(X_i)}{X_i}\right] = \frac{C'(Q)}{P} \varepsilon_{Q,X_i}. \tag{9}$$

Hence the ratio estimator using the revenue elasticity recovers

$$\hat{\mu}_R = \frac{\varepsilon_{R,X_i}}{s_{R,X_i}} = 1 + \frac{\kappa_i'(X_i)}{X_i},$$

and the ratio estimator using the output elasticity recovers

$$\hat{\mu}_Q = \frac{\varepsilon_{Q,X_i}}{s_{R,X_i}} = \mu\left[1 + \frac{\kappa_i'(X_i)}{X_i}\right].$$

Why might it be more common to estimate $\hat{\mu}_R > 1$ than $\hat{\mu}_R < 1$ when using firm-level data? One hypothesis is that adjustment costs are asymmetrical. It is less costly to use less of an input than previously planned than to use more of an input. If this is the case then on average we would recover $\hat{\mu}_R > 1$. Similarly if firms are growing on average we would recover $\hat{\mu}_R > 1$ on average.

The argument above effectively assumes that observed input costs are $W_i X_i$ rather than $W_i X_i + W_i \kappa_i(X_i)$. If this is the measure of observed input costs then

$$s_{R,X_i} = \frac{W_i X_i + W_i \kappa_i(X_i)}{PQ}$$

and we obtain

$$\frac{W_i X_i + W_i \kappa_i'(X_i)}{PQ} = \frac{\lambda}{P} \varepsilon_{Q,X_i}$$
$$\hat{\mu}_Q = \frac{\varepsilon_{Q,X_i}}{s_{R,X_i}} = \mu \left(\frac{X_i + \kappa_i'(X_i)}{X_i + \kappa_i(X_i)} \right)$$

so wedge > 1 whenever $\kappa' > \kappa$.

Neither of the two cases that are typically considered in the literature lead to a bias. The variable input case is $\kappa_i = 0$, in which case the bias disappears. The fixed input case is one in which $X_i \to \overline{X_i}$ in which case the bias also disappears. (Note, however that the fixed input case is not the limit as $\kappa_i \to \infty$, and so is not a special case of the model with adjustment cost model. When $\kappa_i \to \infty$ in the adjustment cost model, the bias remains even in the limit, even though $X_i \to \overline{X_i}$).

B Inputs that influence demand

In this section we show that even if output elasticities are available, markup estimates are biased whenever the variable factor of production is used partly to affect demand in addition to producing output.

We assume that the firm's production function is as in Section 2, but that its revenue is now given by

$$R = P(Q, D) Q$$

where D is a demand shifter. The firm can influence the level of demand through the use of inputs according to the function

$$D = F_D(X_{1D}, X_{2D}, \dots).$$

We denote the amount of input i used in production as X_{iQ} and the amount used in influencing demand as X_{iD}. The total quantity of input i used by the firm is $X_i = X_{iD} + X_{iQ}$.

The profit maximization problem of the firm is now

$$\Pi = \max_{Q,D} P(Q, D) Q - C_Q(Q) - C_D(D) \tag{10}$$

where $C_Q(Q)$ is the firm's cost function for producing output, defined by

$$C_Q(Q) := \min_{X_{iY}} \sum_i X_{iY} W_i \tag{11}$$
$$\text{subject to}$$
$$Q \leq F_Q(X_{1Q}, X_{2Q}, \dots)$$

and $C_D(D)$ is the firm's cost function for influencing demand, defined by

$$C_D(D) := \min_{X_{iD}} \sum_i X_{iD} W_i \tag{12}$$
$$\text{subject to}$$
$$D \leq F_D(X_{1D}, X_{2D}, \dots)$$

The optimality conditions from the profit maximization problem (10) are

$$\varepsilon_{P,Q} + 1 = \frac{C'_Q(Q)}{P} \tag{13}$$

$$\varepsilon_{P,D} = \frac{C'_D(D)\,D}{PQ} \tag{14}$$

where $\varepsilon_{P,D}$ describes the effect of the demand shifter on the price that a firm can charge for a given quantity of output. As in the previous section, the optimal markup of price over marginal production cost is $\mu := \left[\frac{C'_Q(Q)}{P}\right]^{-1} = (1+\varepsilon_{P,Q})^{-1}$.

The FOC for the production cost minimization problem (11) yields the relationship

$$s_{R,X_{iQ}} = \frac{C'_Q(Q)}{P}\varepsilon_{Q,X_{iQ}} \tag{15}$$

where $s_{R,X_{iQ}}$ is the share of revenue paid to input i for use in producing output, and $\varepsilon_{Q,X_{iQ}}$ is the elasticity of output to the use of input i for production. It follows from equation (15) that if one could observe X_{iQ} separately from X_i then the ratio estimator would correctly recover the markup.

However, in practice we observe only the total usage of an input $X_i = X_{iQ} + X_{iD}$, rather then the usage in different activities separately. Using the FOC for the cost minimization problem for influencing demand (12) yields the relationship

$$s_{R,X_{iD}} = \frac{C'_D(D)\,D}{PQ}\varepsilon_{D,X_{iD}} \tag{16}$$

Combining (13),(14), (15) and (16) yields an expression for the total revenue share of input X_i

$$s_{R,X_i} = (1+\varepsilon_{P,Q})\,\varepsilon_{Q,X_{iQ}} + \varepsilon_{P,D}\varepsilon_{D,X_{iD}} \tag{17}$$

To see what the ratio estimator recovers, note that the optimality condition for allocating an input X_i between producing goods X_{iQ} and influencing demand X_{iD} implies

$$\varepsilon_{Q,X_i} = \varepsilon_{Q,X_{iQ}}\varepsilon_{X_{iQ},X_i} + \varepsilon_{Q,X_{iD}}\varepsilon_{X_{iD},X_i} = \varepsilon_{Q,X_{iQ}}\varepsilon_{X_{iQ},X_i} \tag{18}$$

This means that in order to correctly recover the output elasticity of an input X_i, it is necessary to separately observe the part of that input that is actually used in producing goods as long as $\varepsilon_{X_{iQ},X_i} \neq 1$. The fact that a firm uses inputs partly to influence demand introduces a bias into the estimate of the output elasticity. It also introduces a bias into the estimate of the markup. Combining (17) and (18) reveals that the ratio estimator is given by

$$\hat{\mu}_Q = \mu\frac{\varepsilon_{X_{iQ},X_i}}{1+\frac{X_{iD}}{X_{iQ}}}$$

There are however special cases in which $\varepsilon_{X_{iQ},X_i} = 1$, i.e. the share of X_i in production and in influencing demand does not depend on the level of X_i. For example it is sufficient that the firm faces an isoelastic demand curve and F_Q and F_D are Cobb-Douglas. If this is the case, there is no bias the estimate

of the output elasticity, but the ratio estimator is still biased. [16]

$$\hat{\mu_Q} = \mu \frac{1}{1 + \frac{X_{iD}}{X_{iQ}}}.$$

So if the variable input is only used for production and not to influence demand ($X_{iD} = 0$) then the ratio estimator recovers the markup. But if some of the input is used to influence demand, and this component is not separated out, then the ratio estimator will be biased downward. If, over time, the input X_i is increasingly being used to influence demand, then the ratio estimator will fall over time, without any change in the true markup.

Casual observation suggests that at least some part of the workforce currently employed in the corporate sector devotes its energy to influencing demand rather than to producing goods. This suggests that using labor as an input for estimating markups will yield estimates that are hard to interpret. When using the ratio estimator, heterogeneity across firms and industries in the extent to which they use labor for production versus marketing and sales-related expenses will thus manifest as heterogeneity in measured markups.

These observations also help shed light on the difference in the trend in markups that one obtains from Compustat data on US firms when one uses only COGS versus when one includes SGA as the variable input (De Loecker et al. (2020); Traina (2018); De Loecker and Eeckhout (2018)). It seems reasonable to assume that in the COGS bundle, a larger fraction of the inputs is used to produce output and a smaller fraction is used to influence demand, than in the SGA bundle. Thus the downward bias in the ratio estimator is likely to be larger when including SGA in the bundle of variable inputs, versus when using only COGS. Since the cost share of SGA in total revenue has been increasing relative to the cost share of COGS in total revenue, this will manifest as a widening gap between the ratio estimator that uses only COGS and the ratio estimator that also includes SGA. This is precisely what the literature has found.

So far in this section we have proceeded as if output were observed. If only revenue were observed, as in Section 2.1, then the ratio estimator again recovers $\hat{\mu}_R = 1$, regardless of whether the input is being used for production or to influence demand. Given that Compustat data contains only revenue, not output, the aforementioned discussion is relevant only if one believes that the procedures in those papers do successfully recover output elasticities, which we believe they do not.

C Optimal input demand functions

This appendix supplies the derivation of the optimal input demand equation for intermediate inputs under two technology specifications. Section C.1 provides the derivation for a Cobb-Douglas technology and Section C.2 provides that for a nonparametric technology.

C.1 Cobb-Douglas

The three-factor Cobb-Douglas production function for gross output Q_{it}, with Hicks-neutral productivity ω_{it}, is

$$Q_{it} = K_{it}^{\beta_K} L_{it}^{\beta_L} M_{it}^{\beta_M} \exp(\omega_{it})$$

[16] This result does not require that X_{iD} and X_{iQ} are perfect substitutes, but it does require that they satisfy $X_i = f(X_{iD}, X_{iQ})$ where f is a constant-returns-to-scale function. Thanks to Agustin Gutierrez for pointing this out.

Since M_{it} is the single flexible input, the cost minimizing input demand for M_{it} can be obtained by rearranging the Cobb-Douglas production function.

$$M_{it}^{*}=\mathbb{M}\left(K_{it},L_{it},Q_{it}^{*},\omega_{it}\right)=K_{it}^{-\frac{\beta_K}{\beta_M}}L_{it}^{-\frac{\beta_L}{\beta_M}}\left(Q_{it}^{*}\right)^{\frac{1}{\beta_M}}\exp\left(-\frac{1}{\beta_M}\omega_{it}\right) \tag{19}$$

where Q_{it}^{*} is the optimal output level that is taken as given in cost minimization. Then, the minimized total variable cost function is

$$\mathbb{C}\left(K_{it},L_{it},P_{it}^{M},Q_{it}^{*},\omega_{it}\right)\equiv P_{it}^{M}\mathbb{M}\left(K_{it},L_{it},Q_{it}^{*},\omega_{it}\right) \tag{20}$$

where P_{it}^{M} is the unit input price of M_{it} that firm i takes as given. Taking the demand system $P_{it}=P_t\left(Q_{it}\right)$, where $P_t'\left(Q_{it}\right)\leq 0$, and the total cost function $\mathbb{C}\left(K_{it},L_{it},P_{it}^{M},Q_{it}^{*},\omega_{it}\right)$ as given, firm i chooses Q_{it} to maximize its static profits.

$$\max_{Q_{it}}\left\{P_t\left(Q_{it}\right)Q_{it}-\mathbb{C}\left(K_{it},L_{it},P_{it}^{M},Q_{it},\omega_{it}\right)\right\}$$

The first order condition in profit maximization equates marginal revenue to marginal cost.

$$P_t\left(Q_{it}^{*}\right)\left(\frac{\varepsilon_{P,Q}\left(Q_{it}^{*}\right)-1}{\varepsilon_{P,Q}\left(Q_{it}^{*}\right)}\right)=P_{it}^{M}\frac{\partial\mathbb{M}\left(K_{it},L_{it},Q_{it}^{*},\omega_{it}\right)}{\partial Q_{it}^{*}} \tag{21}$$

where $\varepsilon_{P,Q}\left(Q_{it}\right)$ is the price elasticity of demand defined as

$$\varepsilon_{P,Q}\left(Q_{it}\right)\equiv-\frac{P_t\left(Q_{it}\right)}{P_t'\left(Q_{it}\right)Q_{it}}$$

Equation (21) identifies the optimal markup function $\mu_{it}^{*}=\mu_t\left(Q_{it}^{*}\right)$ under monopolistic competition in terms of the demand elasticity.

$$\mu_t\left(Q_{it}^{*}\right)\equiv P_t\left(Q_{it}^{*}\right)\left(P_{it}^{M}\frac{\partial\mathbb{M}\left(K_{it},L_{it},Q_{it}^{*},\omega_{it}\right)}{\partial Q_{it}^{*}}\right)^{-1}=\frac{\varepsilon_{P,Q}\left(Q_{it}^{*}\right)}{\varepsilon_{P,Q}\left(Q_{it}^{*}\right)-1}$$

Applying the functional form in equation (19) to the FOC in equation (21) and solving for $q_{it}^{*}=\ln Q_{it}^{*}$ gives

$$q_{it}^{*}=\frac{\beta_M}{1-\beta_M}\ln\beta_M+\frac{\beta_k}{1-\beta_M}k_{it}+\frac{\beta_l}{1-\beta_M}l_{it}+\frac{\beta_M}{1-\beta_M}\left(p_{it}^{*}-\ln\mu_{it}^{*}-p_{it}^{M}\right)+\frac{1}{1-\beta_M}\omega_{it} \tag{22}$$

where $p_{it}^{M}\equiv\ln P_{it}^{M}$ and $p_{it}^{*}\equiv\ln P_t\left(Q_{it}^{*}\right)$. Using equation (22) to substitute for q_{it}^{*} in equation (19) produces the desired micro-founded optimal input demand equation for m_{it} in terms of the state variables $(k_{it},l_{it},\omega_{it})$, the exogenous input price p_{it}^{M}, and the endogenous optimal output price p_{it}^{*} and markup μ_{it}^{*}.

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

C.2 Nonparametric technology

The nonparametric three-factor production function for gross output with productivity ω_{it} is

$$Q_{it}=F_t\left(K_{it},L_{it},M_{it},\omega_{it}\right) \tag{23}$$

The only restriction we impose on the function $F_t\left(\cdot\right)$ is that it is continuous and twice differentiable with respect to its arguments. We index the function $F_t\left(\cdot\right)$ with a subscript t to allow for technological change

over time. As in Section C.1, M_{it} is the single flexible input. Inverting equation (23) produces the cost-minimizing input demand for M_{it}.

$$M_{it}^{*} = F_{t}^{-1}\left(K_{it}, L_{it}, Q_{it}^{*}, \omega_{it}\right) \tag{24}$$

The minimized total variable cost function is

$$\mathbb{C}_{t}\left(K_{it}, L_{it}, P_{it}^{M}, Q_{it}^{*}, \omega_{it}\right) \equiv P_{it}^{M} F_{t}^{-1}\left(K_{it}, L_{it}, Q_{it}^{*}, \omega_{it}\right)$$

Given a demand system $P_{it} = P_{t}\left(Q_{it}\right)$, the first order condition in profit maximization is

$$\frac{P_{it}^{*}}{\mu_{it}^{*}} = P_{it}^{M} \frac{\partial F_{t}^{-1}\left(K_{it}, L_{it}, Q_{it}^{*}, \omega_{it}\right)}{\partial Q_{it}^{*}} \tag{25}$$

Given a functional form for $F_{t}\left(\cdot\right)$, equation (25) can be solved for the optimal output level Q_{it}^{*}.

$$Q_{it}^{*} = \mathbb{Q}_{t}\left(K_{it}, L_{it}, P_{it}^{M}, \omega_{it}, P_{it}^{*}, \mu_{it}^{*}\right) \tag{26}$$

Using equation (26) to substitute for Q_{it}^{*} in equation (24) yields the micro-founded optimal input demand function for intermediate inputs.

$$\begin{aligned} M_{it}^{*} &= F_{t}^{-1}\left(K_{it}, L_{it}, \mathbb{Q}_{t}\left(K_{it}, L_{it}, P_{it}^{M}, \omega_{it}, P_{it}^{*}, \mu_{it}^{*}\right), \omega_{it}\right) \\ &\equiv \mathbb{M}_{t}\left(K_{it}, L_{it}, P_{it}^{M}, \omega_{it}, P_{it}^{*}, \mu_{it}^{*}\right) \end{aligned}$$

In the absence of price data on inputs and outputs, the scalar unobservables in the input demand function $\mathbb{M}_{t}\left(\cdot\right)$ are $\left(P_{it}^{M}, \omega_{it}, P_{it}, \mu_{it}\right)$.

D Learning about variation in markups from variation in the cost share only

Without a way to estimate the output elasticity for a flexible input consistently from typical production data, we cannot use the ratio estimator to learn about the level of price-cost markups. We can however still use insights from the production approach to learn about variation in markups across firms. This variation can be studied using a regression model for the log of the cost share in total revenue for a perfectly flexible input. We sketch this 'cost share approach' to studying markups in this appendix.

As discussed in Section 2, the ratio estimator relies on the relationship $\mu = \frac{\varepsilon_{Q,X_i}}{s_{R,X_i}}$. Taking logs and rearranging, we obviously have $-\ln s_{R,X_i} = -\ln \varepsilon_{Q,X_i} + \ln \mu$. First consider the three factor, Cobb-Douglas case in which intermediate inputs (M) is the perfectly flexible input, as discussed in Section 3. Here $\ln s_{R,M} = (p^{M} + m) - (p + q)$ is the log of the true cost share in revenue for intermediate inputs, and $\ln \varepsilon_{Q,M} = \ln \beta_{M}$ is a constant term. Letting $\ln s_{it} = (p_{it}^{M} + m_{it}) - (p_{it} + y_{it})$ denote the log of the observed cost share in revenue for firm i in period t, we then have

$$-\ln s_{it} = -\ln \beta_{M} + \ln \mu_{it} + \varepsilon_{it} \tag{27}$$

where $y_{it} = q_{it} + \varepsilon_{it}$ as before.[17]

Without a consistent estimate of the output elasticity (β_{M}), it is clear that the mean of the log of the observed cost shares conflates the log of the output elasticity and the mean of the log of the price-cost

[17] For simplicity, we assume here that this is the only source of measurement error in the log of the observed cost share in revenue. In the Cobb-Douglas case, we can easily allow for (multiplicative) measurement error in both the numerator and the denominator of the cost share for intermediate inputs.

markups, and does not separately identify the latter. Nevertheless, under the maintained assumption that the output elasticity is common to all the firm-year observations, we can use this relation to study variation in price-cost markups. For example, if the binary dummy D_{it} indicates whether or not firm i in period t is an exporter, we can specify a linear relationship between log markups and export status

$$\ln \mu_{it} = \delta_0 + \delta_1 D_{it} + \nu_{it} \tag{28}$$

as in De Loecker and Warzynski (2012). Substituting (28) into (27), we have the linear specification

$$-\ln s_{it} = (\delta_0 - \ln \beta_M) + \delta_1 D_{it} + (\varepsilon_{it} + \nu_{it}). \tag{29}$$

In the Cobb-Douglas case, we can thus learn about the *association* between log markups and export status from a simple regression of the log of the observed cost share in revenue for a flexible input on a constant and the export status dummy.[18]

For more general Hicks-neutral gross output production functions, we can write the log of the output elasticity $\ln \varepsilon_{Q,M} = f(k, l, m)$,[19] in which case (29) becomes

$$-\ln s_{it} = g(k_{it}, l_{it}, m_{it}) + \delta_1 D_{it} + (\varepsilon_{it} + \nu_{it}) \tag{30}$$

where $g(k_{it}, l_{it}, m_{it}) = \delta_0 - f(k_{it}, l_{it}, m_{it})$. We can then learn about the association between log markups and export status either by approximating $g(k_{it}, l_{it}, m_{it})$ using a flexible functional form, or by estimating (30) using semi-parametric methods for partially linear models (Robinson (1988)).

This cost share approach allows us to learn about some forms of variation across firms in markups under essentially the same assumptions needed for the production approach, but without requiring a consistent estimate of the output elasticity. Except in the Cobb-Douglas case, we could not use this approach to study the association between markups and measures of firm size (e.g. the log of employment, l_{it}) or measures of factor intensity (e.g. the log of the capital-labor ratio, $k_{it} - l_{it}$); we may also have low power to detect significant association between markups and observed firm characteristics that are strongly correlated with functions of the production inputs. In principle, this approach could also be used to study trends in markups over time, as in De Loecker et al. (2020). However, it should be emphasized that the trend in the log of the cost share in revenue for a flexible input identifies the trend in the log of the markup only under the maintained assumption that the output elasticity is stable over time, which cannot be verified without a way of estimating the output elasticity consistently for different sub-periods.

[18] As in De Loecker and Warzynski (2012), additional controls can be included in this regression specification, but OLS is still unlikely to consistently estimate the causal effect of exporting on markups. If the sample used to estimate (29) pools data for firms in several sectors, sector dummies can be used to allow for heterogeneity in the output elasticity β_M between sectors.

[19] For example, in the translog case, we have $f(k, l, m) = \ln(\beta_M + \beta_{KM} k + \beta_{LM} l + \beta_{MM} m)$.