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

Steve Bond† Arshia Hashemi‡ Greg Kaplan§ Piotr Zoch¶

December 2, 2020

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) without seperate data on output prices and quantities, as is typically the case, the output elasticity is not identified non-parametrically from estimation of the revenue production function; (ii) even with data on output prices and quantities, it is challenging 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.

JEL Codes: D2, D4, L1, L2

Keywords: Markups, Output Elasticity, Revenue Elasticity, Production Functions.

*We thank Tugce Turk for excellent research assistance. We thank Daniel Ackerberg, Susanto Basu, Jan De Loecker, Jordi Gali, Chad Syverson and Ali Hortacsu for helpful comments and suggestions.

†University of Oxford; steve.bond@nuffield.ox.ac.uk

‡University of Chicago; arshiahashemi@uchicago.edu

§University of Chicago and NBER; gkaplan@uchicago.edu

¶University of Chicago; 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 of the markup from variants of the ratio estimator. 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 3.1, 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. 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. We also show that even if separate data on prices and quantities are available, it is still challenging to obtain consistent estimates of output elasticities for flexible inputs, particularly if only a firm-level price *index* is available. In Section 3.2 we discuss the possibilities for estimating the revenue elasticity consistently with data on revenues, 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 and prices 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 **??** 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 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.

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.[3] 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.[4]

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 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 also violates the scalar unobservability condition needed to obtain a valid control function for unobserved total factor productivity in the first stage of the two

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

[4] This point has also been made in a recent paper by Doraszelski and Jaumandreu (2019).

stage estimation procedures that are often used in this setting. 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

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

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}.[5] 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 $\varepsilon_{QMit} \coloneqq \partial q_{it}/\partial m_{it}$.

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 $r_{it} \coloneqq p_{it} + q_{it}$.[6] To analyze this further, we assume that each firm faces a downward-sloping inverse demand curve of the form

$$p_{it} = p(q_{it}, \xi_{it}) \tag{7}$$

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

$$r_{it} = (p_{it} + q_{it}) = f(k_{it}, l_{it}, m_{it}) + (p_{it} + \omega_{it}) \tag{8}$$

[5] 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.

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

The dependence of intermediate inputs (m_{it}) on unobserved total factor productivity (ω_{it}) raises issues for the consistent estimation of the output elasticity ε_{QMit} from the quantity production function (6) 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 (8) raises more fundamental issues when firms have market power, and their output price depends on q_{it} from (7), 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'. 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 ε_{QMit} is not identified non-parametrically from estimation of the revenue production function (8).

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 (7). 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 (8). 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}) on 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 (8). However, the special cases with zero pass through of demand shocks on to the output price require strong parametric restrictions on the form of both the quantity production function (6) and the inverse demand curve (7), 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

$$\Pi(k_{it}, l_{it}; \omega_{it}, \xi_{it}, p^M_{it}) \coloneqq P_{it}(Q_{it})Q_{it} - P^M_{it} M_{it} \tag{9}$$

where P^M_{it} is the price of one unit of intermediate inputs for firm i in period t, p^M_{it} is the log of this price, and we assume that the firm takes this input price as given; the input price is observed by the firm, and may be observed or unobserved by the researcher.

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 $(\omega_{it}, \xi_{it}, p^M_{it})$:

$$\begin{aligned} m_{it} &= g_m(k_{it}, l_{it}; \omega_{it}, \xi_{it}, p^M_{it}) \\ p_{it} &= g_p(k_{it}, l_{it}; \omega_{it}, \xi_{it}, p^M_{it}) \end{aligned} \tag{10}$$

These decision rules then indicate that any informative instrument for m_{it} in (8) 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 (8) 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}.[7]

In this context, variation in the input price p^M_{it} 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^M_{it} from the decision rule for the output price in (10). The demand 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 on 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 demand curve for each firm.[8] 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 (8), provided they are also uncorrelated with ω_{it}. This *requires* heterogeneity across firms in the inverse demand curves, and the output elasticity for the flexible input

[7] 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.

[8] 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 $\beta_M = \varepsilon_{QM} = 1$ 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 $\eta^{-1} = -\varepsilon_{PQ} > 0$.

is identified *only* at one point ($\varepsilon_{QM} = 1$) 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.

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 C 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

$$y_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + (\omega_{it} + \varepsilon_{it}) \tag{11}$$

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.[9] 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 $\beta_M = \varepsilon_{QM}$.

We again assume that the firm chooses the level of intermediate inputs to maximize net revenue in (9), taking the input price as given. Without specifying the form of the inverse

[9] An alternative interpretation of the two error components in (11) 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.

demand curve (7), we show in Appendix C that the optimal choice of intermediate inputs satisfies the first order condition

$$m_{it} = \frac{\ln \beta_M}{1-\beta_M} + \left(\frac{\beta_K}{1-\beta_M}\right) k_{it} + \left(\frac{\beta_L}{1-\beta_M}\right) l_{it} + \left(\frac{1}{1-\beta_M}\right) \left(p_{it} - \ln \mu_{it} - p_{it}^M + \omega_{it}\right) \tag{12}$$

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 (11). If the researcher has data on the input price p_{it}^M, 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 (11). In that case we have the additional moment condition $\mathrm{E}[p_{it}^M u_{it}] = 0$, and the parameter vector $(\beta_K, \beta_L, \beta_M)$ is identified from the estimation of the quantity production function (11).

If the researcher does not have data on the price of the flexible input, 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_{it}^M 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_{it}^M or z_{it}, the first order condition (12) implies that the lagged input $m_{i,t-1}$ provides a valid and informative instrument for the explanatory variable m_{it} in (11), and in this case we have the additional (informative) moment condition $\mathrm{E}[m_{i,t-1} u_{it}] = 0$.[10]

For price-taking firms, it is well known that identification of the output elasticity for a

[10] 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 (11), 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}.

flexible input from estimation of the quantity production function requires variation across firms in the price of the flexible input.[11] 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 processes

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

$$\omega_{it} = \rho\omega_{i,t-1} + \upsilon_{it} \tag{13}$$

with $|\rho| < 1$, in which the productivity innovations υ_{it} are independent across firms and serially uncorrelated. Substituting for ω_{it} in (13) from (11), and similarly for $\omega_{i,t-1}$, results in a quasi-differenced representation of the quantity production function in which the error term is now $u_{it} \coloneqq \upsilon_{it} + \varepsilon_{it} - \rho\varepsilon_{i,t-1}$. 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, 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 $s \leqslant t-2$.[12] 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 processes

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

[12] This restriction follows naturally if the ε_{it} component of the error term in (11) is interpreted as a shock to productivity that is not known by the firm when making input decisions in period t.

The same approach could be used 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 (11). Otherwise, if we replace the linear function $\rho\omega_{i,t-1}$ on the right-hand side of (13) by a non-linear function $\phi(\omega_{i,t-1})$, the presence of the unobserved $\varepsilon_{i,t-1}$ inside this non-linear function when we substitute for $\omega_{i,t-1}$ using (11) will invalidate moment conditions of the kind considered in the previous sub-section.

Here 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 first need to eliminate the measurement error component ε_{it} from the error term of the quantity production function (11), before we substitute for $\omega_{i,t-1}$ in the non-linear function $\phi(\omega_{i,t-1})$.

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.[13] However, there seem to be 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 (11) as a function of observed variables. 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.

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 can express the optimal choice of the flexible input as a function $m_{it} = m(k_{it}, l_{it}, z_{it}, p^M_{it}, \omega_{it})$, as illustrated in (12) for the Cobb-Douglas gross output production function. 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. In this case we can express $z_{it} = z(k_{it}, l_{it}, p^M_{it}, d_t, \omega_{it})$ and hence we can express m_{it} as a function of the observed predetermined inputs, the price of the flexible input, time dummies, and the

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

scalar unobservable ω_{it}. We can invert this function to obtain 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 z_{it} will additionally depend on the unobserved component of the demand shifter, and we can no longer express m_{it} as a function of observed variables and the scalar unobservable ω_{it}. We can still invert the function $m_{it} = m(k_{it}, l_{it}, z_{it}, p^M_{it}, \omega_{it})$ to obtain $\omega_{it} = h(k_{it}, l_{it}, m_{it}, z_{it}, p^M_{it})$, but to make use of this control function in the first stage regression the researcher would need to be able to control for variation in the log of marginal cost z_{it}.[14] 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.[15]

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

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

[15] 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.

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.

This deflated measure of revenue measures the true level of output Q_{it} up to the unknown multiplicative 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 (11)

$$(p_{i0} + y_{it}) = p_{i0} + \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + (\omega_{it} + \varepsilon_{it}) \tag{14}$$

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).[16]

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 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} typically rule out unobserved firm-specific fixed effects. Constructing a valid control function in the first stage specification requires that we can express m_{it} as a function of the scalar unobservable component ω_{it}, and this condition is violated if we have an additional source of unobserved heterogeneity. Consistent estimation of the first stage specification using least squares methods further requires that the included explanatory variables are uncorrelated with the remaining error term after eliminating the ω_{it} component. In the Cobb-Douglas special case, it may be possible to construct a valid control function for $\Delta\omega_{it}$

[16] 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 variation across firms in the level of the input price.

and then apply a two stage procedure to equations in first-differences. Given the highly persistent nature of most data on inputs and output, relying exclusively on equations in first-differences may not be attractive. More generally, it is not clear how to allow for firm-specific intercepts in the estimation of production functions if we wish to allow for non-linear dynamics in the total factor productivity component of the error term.

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. The availability of firm-level data on output prices is thus fundamental to obtaining credible estimates of output elasticities from the estimation of a production function. Even with output price data for individual firms, the output elasticity for a flexible input may not be identified if there is unobserved heterogeneity across firms in the demand schedules, or if only firm-specific output price indices are available, and the log of unobserved total factor productivity follows a non-linear dynamic process. 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 (11) and a Constant Elasticity of Substitution inverse demand curve

$$p_{it} = \delta_t - \eta^{-1} q_{it} + \zeta_{it} \tag{15}$$

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, i.e. we have $\eta = -\varepsilon_{PQ}^{-1}$. The revenue production function in this case is

$$r_{it}^o = (p_{it} + y_{it}) = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + (p_{it} + \omega_{it} + \varepsilon_{it}) \tag{16}$$

where the log of observed revenue $r_{it}^o \coloneqq r_{it} + \varepsilon_{it}$ 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 (16) from the inverse demand curve (15), we obtain the log-linear equation

$$r_{it}^o = \delta_t + \left(\frac{\beta_K}{\mu}\right) k_{it} + \left(\frac{\beta_L}{\mu}\right) l_{it} + \left(\frac{\beta_M}{\mu}\right) m_{it} + \left[\left(\frac{1}{\mu}\right) \omega_{it} + \varepsilon_{it} + \zeta_{it}\right] \tag{17}$$

which relates observed revenue to the observed inputs. Here $\mu = \left(1 - \frac{1}{\eta}\right)^{-1} > 1$ is the markup of price over marginal cost, 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 (17) 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 the three components of the error term $u_{it} \coloneqq \left(\frac{1}{\mu}\right)\omega_{it} + \varepsilon_{it} + \zeta_{it}$ are assumed to be serially uncorrelated, we have moment conditions $\mathrm{E}[(k_{it}, l_{it})u_{it}] = 0$. With persistent variation across firms in the input price p_{it}^{M}, the lagged input $m_{i,t-1}$ provides a valid and informative instrument for m_{it}, and we have the additional (informative) moment condition $\mathrm{E}[m_{i,t-1}u_{it}] = 0$. 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).

In cases where we can estimate these revenue elasticity parameters consistently, we could investigate heterogeneity in the markup parameter μ across sub-samples of firms by including suitable interaction terms in (17), under the maintained assumption that the output elasticities are common to these sub-samples.[17]

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.[18] 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; for example, if both ω_{it} and ζ_{it} follow AR(1) processes, as in (13), but with different autoregressive parameters (ρ). Consistent estimation may also be difficult if the markup parameter μ is not common within 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 (17).[19] Finally, consistent estimation of the revenue elasticities

[17] 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).

[18] 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 (12).

[19] 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 $e_{it} \coloneqq u_{it} + (\beta_i - \beta)x_{it}$. 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.

is likely to be more difficult if the gross output production function and inverse demand curve do not take the convenient log-linear forms that we have considered in this section.

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, *135* (2), 561–644.

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

Forlani, E., R. Martin, G. Mion, and M. Muls, "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

$$\begin{aligned} C(Q) := \min_{X_i} & \sum_i X_i W_i + \sum_i \kappa_i(X_i) W_i \\ & \text{subject to} \\ Q \leq & F_Q(X_1, X_2, \dots) \end{aligned} \tag{18}$$

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

$$s_{R,X_i}\left[1 + \frac{\kappa_i'(X_i)}{X_i}\right] = \frac{C'(Q)}{P} \varepsilon_{Q,X_i}. \tag{19}$$

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

$$\Pi = \max_{Q,D} P(Q, D)\, Q - C_Q(Q) - C_D(D) \tag{20}$$

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

$$C_Q(Q) := \min_{X_{iY}} \sum_i X_{iY} W_i \tag{21}$$
$$\text{subject to}$$
$$Q \leq F_Q(X_{1Q}, X_{2Q}, \dots)$$

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

$$C_D(D) := \min_{X_{iD}} \sum_i X_{iD} W_i \tag{22}$$
$$\text{subject to}$$
$$D \leq F_D(X_{1D}, X_{2D}, \dots)$$

The optimality conditions from the profit maximization problem (20) are

$$\varepsilon_{P,Q} + 1 = \frac{C'_Q(Q)}{P} \tag{23}$$

$$\varepsilon_{P,D} = \frac{C'_D(D)\,D}{PQ} \tag{24}$$

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 (21) yields the relationship

$$s_{R,X_{iQ}} = \frac{C'_Q(Q)}{P}\varepsilon_{Q,X_{iQ}} \tag{25}$$

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 (25) 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 (22) yields the relationship

$$s_{R,X_{iD}} = \frac{C'_D(D)\,D}{PQ}\varepsilon_{D,X_{iD}} \tag{26}$$

Combining (23),(24), (25) and (26) yields an expression for the total revenue share of input X_i

$$s_{R,X_i} = (1 + \varepsilon_{P,Q})\,\varepsilon_{Q,X_{iQ}} + \varepsilon_{P,D}\varepsilon_{D,X_{iD}} \tag{27}$$

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

$$\varepsilon_{Q,X_i} = \varepsilon_{Q,X_{iQ}}\varepsilon_{X_{iQ},X_i} + \varepsilon_{Q,X_{iD}}\varepsilon_{X_{iD},X_i} = \varepsilon_{Q,X_{iQ}}\varepsilon_{X_{iQ},X_i} \tag{28}$$

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 (27) and (28) 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. [20]

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

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

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

[20] 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.

$$M_{it}^{*}=\mathbb{M}\left(K_{it},L_{it},Q_{it}^{*},\omega_{it}\right)=K_{it}^{-\frac{\beta_K}{\beta_M}}L_{it}^{-\frac{\beta_L}{\beta_M}}\left(Q_{it}^{*}\right)^{\frac{1}{\beta_M}}\exp\left(-\frac{1}{\beta_M}\omega_{it}\right) \tag{29}$$

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

$$\mathbb{C}\left(K_{it},L_{it},P_{it}^{M},Q_{it}^{*},\omega_{it}\right)\equiv P_{it}^{M}\mathbb{M}\left(K_{it},L_{it},Q_{it}^{*},\omega_{it}\right) \tag{30}$$

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.

$$P_t\left(Q_{it}^{*}\right)\left(\frac{\varepsilon_{P,Q}\left(Q_{it}^{*}\right)-1}{\varepsilon_{P,Q}\left(Q_{it}^{*}\right)}\right)=P_{it}^{M}\frac{\partial\mathbb{M}\left(K_{it},L_{it},Q_{it}^{*},\omega_{it}\right)}{\partial Q_{it}^{*}} \tag{31}$$

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 (31) 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 (29) to the FOC in equation (31) and solving for $q_{it}^{*}=\ln Q_{it}^{*}$ gives

$$q_{it}^{*}=\frac{\beta_M}{1-\beta_M}\ln\beta_M+\frac{\beta_k}{1-\beta_M}k_{it}+\frac{\beta_l}{1-\beta_M}l_{it}+\frac{\beta_M}{1-\beta_M}\left(p_{it}^{*}-\ln\mu_{it}^{*}-p_{it}^{M}\right)+\frac{1}{1-\beta_M}\omega_{it} \tag{32}$$

where $p_{it}^{M}\equiv\ln P_{it}^{M}$ and $p_{it}^{*}\equiv\ln P_t\left(Q_{it}^{*}\right)$. Using equation (32) to substitute for q_{it}^{*} in equation (29) 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

$$Q_{it}=F_t\left(K_{it},L_{it},M_{it},\omega_{it}\right) \tag{33}$$

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 (33) produces the cost-minimizing input demand for M_{it}.

$$M_{it}^{*} = F_{t}^{-1}\left(K_{it}, L_{it}, Q_{it}^{*}, \omega_{it}\right) \tag{34}$$

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

$$\frac{P_{it}^{*}}{\mu_{it}^{*}} = P_{it}^{M} \frac{\partial F_{t}^{-1}\left(K_{it}, L_{it}, Q_{it}^{*}, \omega_{it}\right)}{\partial Q_{it}^{*}} \tag{35}$$

Given a functional form for $F_{t}\left(\cdot\right)$, equation (35) can be solved for the optimal output level Q_{it}^{*}.

$$Q_{it}^{*} = \mathbb{Q}_{t}\left(K_{it}, L_{it}, P_{it}^{M}, \omega_{it}, P_{it}^{*}, \mu_{it}^{*}\right) \tag{36}$$

Using equation (36) to substitute for Q_{it}^{*} in equation (34) 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

$$-\ln s_{it} = -\ln \beta_{M} + \ln \mu_{it} + \varepsilon_{it} \tag{37}$$

where $y_{it} = q_{it} + \varepsilon_{it}$ as before.[21]

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

[21] 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

$$\ln \mu_{it} = \delta_0 + \delta_1 D_{it} + \nu_{it} \tag{38}$$

as in De Loecker and Warzynski (2012). Substituting (38) into (37), we have the linear specification

$$-\ln s_{it} = (\delta_0 - \ln \beta_M) + \delta_1 D_{it} + (\varepsilon_{it} + \nu_{it}). \tag{39}$$

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.[22]

For more general Hicks-neutral gross output production functions, we can write the log of the output elasticity $\ln \varepsilon_{Q,M} = f(k, l, m)$,[23] in which case (39) becomes

$$-\ln s_{it} = g(k_{it}, l_{it}, m_{it}) + \delta_1 D_{it} + (\varepsilon_{it} + \nu_{it}) \tag{40}$$

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 (40) 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.

[22] 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 (39) pools data for firms in several sectors, sector dummies can be used to allow for heterogeneity in the output elasticity β_M between sectors.

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