Reexamining the De Loecker & Warzynski (2012) method for estimating markups*

Ulrich Doraszelski
University of Pennsylvania†

Jordi Jaumandreu
Boston University‡

December 16, 2021

Abstract

De Loecker & Warzynski (2012) obtain the markup from the firm's cost minimization problem by substituting in estimates of the output elasticity of a variable input and the disturbance that separates actual from planned output. We show, however, that consistently estimating the output elasticity and the disturbance using the procedure developed in Olley & Pakes (1996) and Levinsohn & Petrin (2003) generally requires observing and controlling for the markup. We analyze the resulting biases and discuss alternative approaches to estimation.

*An earlier draft of this paper was circulated in 2020 under the title "The inconsistency of De Loecker & Warzynski's (2012) method to estimate markups and a robust alternative". We are grateful to Ivan Fernandez-Val for useful discussions and to Rigoberto Lopez and Katja Seim for comments. Hyun Soo Suh and Guilherme Neves Silveira provided excellent research assistance.

†Wharton School, Email: doraszelski@wharton.upenn.edu

‡Department of Economics, Email: jordij@bu.edu

1 Introduction

Following a wave of acquisitions in an industry, a policymaker asks an economist if, and by how much, market power has increased. To answer this question, the economist has detailed production data for a sample of firms in the industry at her disposal, including output and input quantities and prices. The economist uses the De Loecker & Warzynski (2012) (henceforth DLW) method to estimate the firm-level markup and then regresses it on a dummy that is one for the acquiring firms in the "after" period and zero otherwise. If the acquisitions have in truth raised the markup of these firms, this regression is likely to tell exactly the opposite. In this paper, we explain what has gone wrong in the exercise of the economist and how to address the problem.

To provide policy advice and answer a variety of empirical and theoretical questions, economists would like to have an easy-to-compute way to estimate the firm-level markup that does not require modelling demand and making detailed assumptions about firm conduct. Bain's (1951) ratio of revenue to variable cost comes close to this ideal but relies on equating inherently unobservable marginal cost with average variable cost.

The production approach to estimating the markup has searched for ways out of this impasse. DLW note that very generally a firm minimizes its cost irrespective of the specifics of demand and firm conduct. They therefore obtain the markup from the FOC for cost minimization by substituting in estimates of the output elasticity of a variable input and the disturbance that separates the firm's actual output as recorded in the data from the output that the firm planned on when it made its input decisions. To obtain the output elasticity and the disturbance separating actual from planned output, DLW use the procedure developed in Olley & Pakes (1996) and Levinsohn & Petrin (2003), implemented as suggested by Ackerberg, Caves & Frazer (2015) (henceforth OP, LP, and ACF), to estimate the production function.

The DLW method has been widely applied to study the distribution of the markup across firms and its evolution over time (see, e.g., De Loecker, Goldberg, Khandelwal & Pavcnik 2016, Brandt, Van Biesebroeck, Wang & Zhang 2017, Brandt, Van Biesebroeck, Wang & Zhang 2019, De Loecker & Scott 2016, De Loecker, Eeckhout & Unger 2020, De Loecker & Eeckhout 2018, Autor, Dorn, Katz, Patterson & Van Reenen 2020).[1] In this paper,

[1] See also Berry, Gaynor & Scott-Morton (2020) for a recent panorama of the industrial organization literature on markups.

we first characterize the circumstances under which the DLW method consistently estimates markups. We then show that outside these circumstances the DLW method produces inconsistent estimates of the output elasticity and the disturbance and therefore biased markups. In particular, the DLW method is not robust to any differences in demand across firms or time unless they are observed by the econometrician in their entirety.

This poses an especially thorny issue because the demand that a firm faces in the output market is the fundamental determinant of the markup that the firm charges. At the same time, the large literatures on demand estimation and productivity analysis make clear that controlling for differences in demand by observables is a daunting task. Papers such as Berry, Levinsohn & Pakes (1995) and Foster, Haltiwanger & Syverson (2008) notably highlight the considerable heterogeneity in demand that remains even after controlling for detailed product attributes or honing in on (nearly) homogenous products. The issue is compounded by the fact that, in imperfectly competitive industries, the demand that a firm faces depends on its rivals, which are partially or completely unobserved in typical production data.

The intuitive reason for the inconsistency of the DLW method is as follows. The OP/LP procedure solves the endogeneity problem in production function estimation by inverting a decision of the firm that the econometrician observes, such as the firm's demand for a variable input, to recover the firm's productivity that the econometrician does not observe. This inversion presumes that two firms that have the same productivity have the same input demand. If there is heterogeneity in demand in addition to heterogeneity in productivity, then this is not the case: two firms that have the same productivity but charge different markups because they face different demands in the output market generally have different input demands. It is therefore no longer possible to express unobserved productivity in terms of observables. Put differently, to use the OP/LP procedure to estimate the production function and obtain the markup, the DLW method would have to observe and control for the markup. In this way, the DLW method is circular.

The observation that the proxy variable methods developed by OP and LP cannot accommodate unobserved demand heterogeneity has been made before. Foster et al. (2008) put it as follows:

> ... idiosyncratic demand shocks make the proxies functions of both technology and demand shocks, thereby inducing a possi-

> ble omitted variable bias. Put simply, proxy methods require a one-to-one mapping between plant-level productivity and the observables used to proxy for productivity. This mapping breaks down if other unobservable plant-level factors besides productivity drive changes in the observable proxy. (p. 403)

At first glance, however, this observation appears irrelevant under the cost minimization assumption of DLW. The purpose of relying on cost minimization instead of profit maximization is precisely to insulate the estimated markup from the specifics of demand and firm conduct. After all, a firm minimizes its cost in most circumstances. De Loecker et al. (2016) accordingly state that (their extension of) the DLW method (to multi-product firms) "does not require assumptions on the market structure or demand curves faced by firms" (p. 445, see also p. 497).

This is an overstatement. We show that in the cost minimization problem the firm's planned output summarizes the demand the firm faces. Because planned output remains unobserved by the econometrician, the firm's cost-minimizing decisions again cannot be inverted to express unobserved productivity in terms of observables. The DLW method therefore either has to rule out any differences in demand across firms or time or assume that they can be fully controlled for by observables.

At a minimum, the conditions required by the DLW method to consistently estimate markups must be justified from a detailed understanding of the market structure and the demands firms face in the industry under study. This negates the purported advantage of relying on cost minimization in the production approach to estimating the markup. Whether one can fully control for any differences in demand across firms or time by observables also remains questionable in light of the large literatures on demand estimation and productivity analysis (Berry et al. 1995, Foster et al. 2008).

We therefore first characterize the bias in the estimates produced by the DLW method that results if there are differences in demand across firms or time that cannot be fully controlled for by observables. We show that the bias permeates the level of the estimated markup and its correlation with variables of interest such as a firm's export status or measures of trade liberalization. Using data from the Spanish manufacturing sector, we then test for the effects of unobserved demand heterogeneity and illustrate their consequences for the estimated markup.

In sum, our paper makes three main contributions. First, we highlight

a not duly appreciated—and sometimes completely overlooked—assumption of the DLW method. OP and LP are careful to rule out unobserved demand heterogeneity.[2] Indeed, this has been codified as the scalar unobservable assumption (Ackerberg, Benkard, Berry & Pakes 2007) in the subsequent literature. DLW take the OP/LP procedure to a different context that allows for heterogeneity in demand and the markup without acknowledging the implication, namely that to use an OP/LP procedure to estimate the production function and obtain the markup, one would have to observe and control for the markup. Second, we contribute by developing the consequences of unobserved demand heterogeneity and characterizing the bias in the estimates produced by the DLW method. Third, we point to dynamic panel methods as alternative approaches to estimation that are robust to any differences in demand across firms or time even if they cannot be fully controlled for by observables.

Our paper is related to Doraszelski & Jaumandreu (2019), where we first note the inconsistency of the DLW method in a setting that emphasizes the implications of biased technological change for markup estimation. More recently, Bond, Hashemi, Kaplan & Zoch (2020) have reiterated our point about the inconsistency of the DLW method, although their focus is on the difficulties for estimating the markup that arise if the econometrician observes revenue rather than the quantity of output.

The remainder of this paper is organized as follows. In Section 2, we recall the setup and the DLW method for estimating the markup. In Section 3, we argue that it is generally not possible to express unobserved productivity in terms of observables. In Section 4, we characterize the bias in the estimates if the economist nevertheless proceeds along the lines of DLW. In Section 5, we provide an empirical application to test for the effects of unobserved demand heterogeneity. We conclude in Section 6 with a discussion of alternative approaches to estimation.

[2]LP assume a perfectly competitive industry where firms act as price takers and thus face the same horizontal demand curve (see p. 322 and Appendix A). OP rule out unobserved demand heterogeneity by assuming that any profitability differences across firms are due to differences in their capital stocks and productivities (see p. 1273). Limiting the state variables in the firm's investment policy to its own capital stock and productivity implicitly abstracts from competition between firms (see also Lemma 3 and Theorem 1 in Pakes (1994)).

2 DLW method

Firm j produces output Q_{jt} in period t with a predetermined amount of capital K_{jt} and freely variable amounts of labor L_{jt} and materials M_{jt}.[3] The production function is

$$Q_{jt} = Q^*_{jt} \exp(\varepsilon_{jt}), \quad Q^*_{jt} = F(K_{jt}, L_{jt}, M_{jt}) \exp(\omega_{jt}), \tag{1}$$

where ω_{jt} is Hicks-neutral productivity that the firm observes before it decides on variable inputs in period t but that remains unobserved by the econometrician. As usual in the literature following OP and LP, ω_{jt} is governed by a first-order Markov process with the law of motion $\omega_{jt} = E(\omega_{jt}|\omega_{jt-1}) + \xi_{jt} = g(\omega_{jt-1}) + \xi_{jt}$, where $g(\omega_{jt-1})$ is expected productivity and ξ_{jt} is the productivity innovation. The disturbance ε_{jt} accounts for the difference between the firm's actual output Q_{jt} as recorded in the data and the output Q^*_{jt} that the firm planned on when it made its input decisions. In contrast to ω_{jt}, ε_{jt} is uncorrelated over time and with the inputs. Because neither the firm nor the econometrician observes ε_{jt}, planned output Q^*_{jt} also remains unobserved by the econometrician.

DLW assume cost minimization in an attempt to avoid specifying demand and firm conduct (pp. 2437–2438 and p. 2443). The firm minimizes variable cost $VC_{jt} = P_{Ljt}L_{jt} + P_{Mjt}M_{jt}$, where P_{Ljt} and P_{Mjt} are the prices of labor and materials, subject to achieving its planned output Q^*_{jt}. The FOC for variable input $X_{jt} \in \{L_{jt}, M_{jt}\}$ is

$$\frac{1}{MC(K_{jt}, P_{Ljt}, P_{Mjt}, Q^*_{jt}, \omega_{jt})} = \frac{\frac{\partial F(K_{jt}, L_{jt}, M_{jt})}{\partial X_{jt}} \exp(\omega_{jt})}{P_{Xjt}}, \tag{2}$$

where the envelope theorem serves to replace the Lagrange multiplier by short-run marginal cost $MC(\cdot)$. As Samuelson (1947) puts it and Hall (1988) first exploits empirically, the FOC says that the marginal productivity of the last dollar must be equal in every use.

The markup is defined as $\mu_{jt} = \frac{P_{jt}}{MC(\cdot)}$, where P_{jt} is the price of output. Rewriting the FOC in equation (2) using the production function in equation

[3] Applications of OP/LP and DLW differ in the identity of the variable input: DLW alternatively assume labor or materials to be freely variable (p. 2457); De Loecker et al. (2016) assume materials to be freely variable (p. 471); and LP assume both labor and materials to be freely variable (p. 322 and p. 339). We adopt the latter assumption merely for concreteness.

(1) yields

$$\mu_{jt} = \frac{P_{jt}Q_{jt}}{P_{Xjt}X_{jt}} \frac{\frac{\partial F(K_{jt},L_{jt},M_{jt})}{\partial X_{jt}} X_{jt}}{F(K_{jt},L_{jt},M_{jt})\exp(\varepsilon_{jt})} = \frac{\beta_X(K_{jt},L_{jt},M_{jt})}{S^R_{Xjt}} \exp(-\varepsilon_{jt}), \quad (3)$$

where $\beta_X(\cdot) = \frac{\partial F(\cdot)}{\partial X_{jt}} \frac{X_{jt}}{F(\cdot)}$ is the output elasticity of variable input X_{jt} and $S^R_{Xjt} = \frac{P_{Xjt}X_{jt}}{P_{jt}Q_{jt}}$ is the expenditure share of the input. Note that S^R_{Xjt} is observed because it is based on actual output Q_{jt} rather than planned output Q^*_{jt}. DLW therefore obtain the markup μ^{DLW}_{jt} of firm j in period t by substituting estimates of $\beta_X(\cdot)$ and ε_{jt} into equation (3).

OP/LP estimation. DLW use the procedure developed by OP and LP, implemented as suggested by ACF, to estimate the output elasticity $\beta_X(\cdot)$ and the disturbance ε_{jt} (pp. 2444–2449). Because actual output $Q_{jt} = Q^*_{jt}\exp(\varepsilon_{jt})$ is observed, estimating the disturbance ε_{jt} is equivalent to estimating planned output Q^*_{jt}.

The OP/LP procedure starts with a function $\omega_{jt} = h(z_{jt})$ that expresses unobserved Hicks-neutral productivity ω_{jt} by a vector of observables z_{jt}. OP use the demand for investment to invert for ω_{jt} and LP the demand for a variable input. z_{jt} correspondingly collects input quantities and prices and all other arguments of the demand that is inverted for ω_{jt}.[4]

Substituting the function $\omega_{jt} = h(z_{jt})$ into equation (1) and taking logs yields

$$q_{jt} = \ln F(K_{jt},L_{jt},M_{jt}) + h(z_{jt}) + \varepsilon_{jt} = \phi(z_{jt}) + \varepsilon_{jt}, \quad (4)$$

where we use lowercase letters to denote logs and $\phi(\cdot)$ is an unknown function that must be estimated nonparametrically. Assuming that ε_{jt} is uncorrelated with z_{jt}[5] and carrying out the regression in equation (4) yields estimates of $\phi(\cdot)$ and the disturbance ε_{jt} that separates actual output q_{jt} from planned output $q^*_{jt} = \phi(z_{jt})$. This is the first step of ACF.

In a second step, the estimate of $\phi(\cdot)$ and the Markovian assumption on Hicks-neutral productivity ω_{jt} serve to estimate the production function and

[4]Following LP, DLW rely on the demand for materials and write $\omega_{it} = h_t(m_{it}, k_{it}, z_{it})$, where i indexes firms and t periods (p. 2446). To economize on the notation, we subsume their m_{it}, k_{it}, and z_{it} into our z_{jt}.

[5]This slightly strengthens the earlier assumption that ε_{jt} is uncorrelated over time and with the inputs.

the implied output elasticity $\beta_X(\cdot) = \frac{\partial \ln F(\cdot)}{\partial x_{jt}}$ by carrying out the regression

$$q_{jt} = \ln F(K_{jt}, L_{jt}, M_{jt}) + g\left(\widehat{\phi}(z_{jt-1}) - \ln F(K_{jt-1}, L_{jt-1}, M_{jt-1})\right) + \xi_{jt} + \varepsilon_{jt}, \tag{5}$$

where the conditional expectation function $g(\cdot)$ is estimated nonparametrically. Note that any variable input that the firm decides on after it observes ω_{jt} is correlated with the productivity innovation ξ_{jt} and must be instrumented for.

In the arguments z_{jt} of the function $\omega_{jt} = h(z_{jt})$, DLW include any "additional variables potentially affecting optimal input demand choice" and advise that "[t]he exact variables to be included ... depend on the application but will definitely capture variables leading to differences in optimal input demand across firms such as input prices" (p. 2446). As we show in Section 3, another variable affecting optimal input demand is planned output Q^*_{jt}. Because planned output Q^*_{jt} is unobserved by the econometrician, it is generally not possible to replace unobserved Hicks-neutral productivity ω_{jt} by observables z_{jt}. The OP/LP procedure therefore cannot be used to estimate the output elasticity $\beta_X(\cdot)$ and the disturbance ε_{jt} (or, equivalently, planned output Q^*_{jt}) to be plugged into equation (3) to obtain the markup. Put differently, to use the OP/LP procedure to estimate planned output Q^*_{jt}, the DLW method would have to observe and control for Q^*_{jt}. In Section 4, we then explore the consequences of ignoring the dependence of the function $\omega_{jt} = h(z_{jt})$ on planned output Q^*_{jt} for the estimated markup.

3 Inverting for unobserved productivity

Inverting a variable input. LP use the demand for a variable input to invert for ω_{jt}. The solution to the cost minimization problem in Section 2 is the variable cost function $VC\left(K_{jt}, P_{Ljt}, P_{Mjt}, \frac{Q^*_{jt}}{\exp(\omega_{jt})}\right)$. From Shephard's lemma, the demand for variable input X_{jt} is thus

$$X_{jt} = \frac{\partial VC\left(K_{jt}, P_{Ljt}, P_{Mjt}, \frac{Q^*_{jt}}{\exp(\omega_{jt})}\right)}{\partial P_{Xjt}}. \tag{6}$$

While this expression can be inverted for $\frac{Q^*_{jt}}{\exp(\omega_{jt})}$, with planned output Q^*_{jt} being unobserved it cannot be inverted for ω_{jt}; hence, it is not possible to

replace unobserved Hicks-neutral productivity ω_{jt} by observables z_{jt}. Combining the demands for two or more variable inputs does not resolve the problem.

Another way to see the problem is to go back to the FOC in equation (2). Because marginal cost is inherently unobservable, the FOC in equation (2) cannot by itself be used to express unobserved Hicks-neutral productivity ω_{jt} in terms of observables z_{jt}. Thus proceeding, marginal cost is related to variable cost as

$$MC(K_{jt}, P_{Ljt}, P_{Mjt}, Q^*_{jt}, \omega_{jt}) = \frac{\partial VC\left(K_{jt}, P_{Ljt}, P_{Mjt}, \frac{Q^*_{jt}}{\exp(\omega_{jt})}\right)}{\partial Q^*_{jt}} \exp(-\omega_{jt}). \tag{7}$$

Substituting into the FOC in equation (2), ω_{jt} cancels and the resulting expression

$$\frac{1}{\frac{\partial VC\left(K_{jt}, P_{Ljt}, P_{Mjt}, \frac{Q^*_{jt}}{\exp(\omega_{jt})}\right)}{\partial Q^*_{jt}}} = \frac{\frac{\partial F(K_{jt}, L_{jt}, M_{jt})}{\partial X_{jt}}}{P_{Xjt}}$$

can again be inverted for $\frac{Q^*_{jt}}{\exp(\omega_{jt})}$ but not for ω_{jt}.

Inverting investment. OP use the demand for investment to invert for ω_{jt}. OP derive the demand for investment from a dynamic profit maximization problem (pp. 1270–1273). This requires OP to take a stand on demand in the output market and firm conduct, which is what DLW intend to avoid. One may alternatively start from a dynamic cost minimization problem (see, e.g., Doraszelski & Jaumandreu 2019), where the firm chooses capital, labor, and materials, possibly subject to adjustment costs, to achieve a sequence of planned outputs Q^*_{jt}. In this case, the demand for investment is a function of $\frac{Q^*_{jt}}{\exp(\omega_{jt})}$ and an analogous problem to the one just reviewed arises.

Controlling for planned output. As we have shown above, as long as planned output Q^*_{jt} is unobserved by the econometrician, it is not possible to replace unobserved Hicks-neutral productivity ω_{jt} by observables z_{jt}. The obvious way around this problem is to assume that planned output Q^*_{jt} can itself be controlled for by a subset of the observables z_{jt}, i.e., that there exists a function $Q^*_{jt} = D(z^D_{jt})$ mapping observables $z^D_{jt} \subseteq z_{jt}$ into planned

output Q^*_{jt}. This is the intuition behind DLW's broad interpretation of z_{jt}. De Loecker et al. (2016) include variables such as location, product dummies, export status, input and output tariffs, market share, and the price of output in z_{jt} (p. 466). Output tariffs, for example, clearly play no role in the cost minimization problem; the only reason to include them in z_{jt} is as an attempt to control for Q^*_{jt}.

The large literatures on demand estimation and productivity analysis cast doubt on any attempt to control for planned output Q^*_{jt} by observables z^D_{jt}. Berry et al. (1995) stress the importance of the unobserved characteristic that remains even after including detailed product attributes in the specification of demand. Foster et al. (2008) similarly highlight the considerable heterogeneity in demand that remains even after honing in on (nearly) homogenous products. Hence, the demand the firm faces is $Q^*_{jt} = D(z^D_{jt}, \delta_{jt})$, where the demand shock δ_{jt} captures unobserved demand heterogeneity in the sense of any differences in demand across firms or time that remain after controlling for observables z^D_{jt}.

Note that under imperfect competition δ_{jt} includes not only the unobserved product characteristic of the firm under consideration but also those of its rivals. Moreover, the demand the firm faces depends on its rivals' prices under Bertrand competition and on its rivals' (planned) quantities under Cournot competition.[6] Changes in firm conduct due to a wave of acquisitions (as in the opening paragraph of Section 1) cause changes in rivals' prices and quantities. To the extent that these variables are partially or completely unobserved, as they are in production data that covers a sample of firms, they become part of δ_{jt}.[7]

In sum, in the cost minimization problem the firm's planned output $Q^*_{jt} = D(z^D_{jt}, \delta_{jt})$ summarizes the demand the firm faces.[8] There is little reason to

[6]Instead of thinking of $Q^*_{jt} = D(z^D_{jt}, \delta_{jt})$ as one of the equations in the demand system for the industry under study, we can think of $D(\cdot)$ as the firm's residual demand in the sense of Baker & Bresnahan (1985). In this case, $D(\cdot)$ encapsulates how the industry equilibrium changes as the focal firm changes its price or quantity. While this obviates accounting for rivals' prices or quantities, $D(\cdot)$ instead depends on assumptions about firm conduct and on rivals' marginal costs and thus their unobserved productivities.

[7]The common practice of letting the function $\omega_{jt} = h(z_{jt})$ vary by period may partly absorb time-series variation but of course not cross-sectional variation due to unobserved differences in demand across firms.

[8]While the firm's cost-minimizing decisions depend indirectly on δ_{jt} through Q^*_{jt}, its profit-maximizing decisions depend directly on δ_{jt} and thus cannot be used either to replace Hicks-neutral productivity ω_{jt} by observables z_{jt}. DLW do not state if the inverse

believe that $\delta_{jt} = 0$ as required by DLW. At the very least, assuming $\delta_{jt} = 0$ requires a careful justification starting from the specification of demand and assumptions on firm conduct, thus negating the purported advantage of the production approach and relying on cost minimization to estimate markups over the demand approach.

In Sections 4 and 5, we further develop the consequences of unobserved demand heterogeneity for the DLW method. We first characterize the bias in the estimated markup resulting from $\delta_{jt} \neq 0$. Then we provide an empirical application to test for the effects of $\delta_{jt} \neq 0$.

4 Bias in estimated markup

If there are differences in demand across firms or time that cannot be fully controlled for by z_{jt}, then $\delta_{jt} \neq 0$ and equation (4) becomes

$$q_{jt} = \phi(z_{jt}, \delta_{jt}) + \varepsilon_{jt}. \tag{8}$$

Regressing q_{jt} on observables z_{jt} in the first step of ACF, however, yields an estimate of the conditional expectation $E(q_{jt}|z_{jt}) = E(\phi(z_{jt}, \delta_{jt})|z_{jt}) = \widetilde{\phi}(z_{jt})$, where the first equality uses that ε_{jt} is uncorrelated with z_{jt}. We thus write the first-stage regression as

$$q_{jt} = \widetilde{\phi}(z_{jt}) + \phi(z_{jt}, \delta_{jt}) - \widetilde{\phi}(z_{jt}) + \varepsilon_{jt} = \widetilde{\phi}(z_{jt}) + \zeta_{jt} + \varepsilon_{jt} = \widetilde{\phi}(z_{jt}) + \widetilde{\varepsilon}_{jt},$$

where the prediction error $\zeta_{jt} = \phi(z_{jt}, \delta_{jt}) - \widetilde{\phi}(z_{jt})$ is mean independent of z_{jt} by construction and $\widetilde{\varepsilon}_{jt} = \zeta_{jt} + \varepsilon_{jt}$ is uncorrelated with z_{jt}.

We develop three alternative characterizations of the prediction error ζ_{jt} that are helpful in assessing the bias in the estimated disturbance in the first step of ACF and the estimated output elasticity in the second step. From the production function in equation (1), we have $\phi(z_{jt}, \delta_{jt}) = \ln F(K_{jt}, L_{jt}, M_{jt}) + \omega_{jt}$. Assuming for simplicity that the demand the firm faces takes the form $Q^*_{jt} = D(z^D_{jt}) \exp(\delta_{jt})$, we also have $\phi(z_{jt}, \delta_{jt}) = \ln D(z^D_{jt}) + \delta_{jt}$. It follows that[9]

$$\zeta_{jt} = \omega_{jt} - E(\omega_{jt}|z_{jt}) = \delta_{jt} - E(\delta_{jt}|z_{jt}). \tag{9}$$

in their equation (9) is derived from the profit-maximizing or cost-minimizing demand for materials.

[9]Recall that by definition $\zeta_{jt} = \phi(z_{jt}, \delta_{jt}) - E(\phi(z_{jt}, \delta_{jt})|z_{jt})$. The first equality in equation (9) follows from $\phi(z_{jt}, \delta_{jt}) = \ln F(K_{jt}, L_{jt}, M_{jt}) + \omega_{jt}$ and the fact that z_{jt} includes input quantities (see Section 2) and the second equality from $\phi(z_{jt}, \delta_{jt}) = \ln D(z^D_{jt}) + \delta_{jt}$ and $z^D_{jt} \subseteq z_{jt}$ (see Section 3).

Our first two characterizations in equation (9) show that ζ_{jt} covaries with any part of Hicks-neutral productivity ω_{jt} and any part of the demand shock δ_{jt} that is not captured by observables z_{jt}.

To develop our third characterization of ζ_{jt}, we use the demand for materials M_{jt} in equation (6) to write the ratio $\frac{Q^*_{jt}}{\exp(\omega_{jt})}$ as

$$\frac{Q^*_{jt}}{\exp(\omega_{jt})} = u(K_{jt}, M_{jt}, P_{Ljt}, P_{Mjt}),$$

where $u(\cdot)$ is an unknown function. Substituting into equation (7), we next write marginal cost MC_{jt} as

$$MC_{jt} = v(K_{jt}, M_{jt}, P_{Ljt}, P_{Mjt}) \exp(-\omega_{jt}),$$

where $v(\cdot)$ is an unknown function. The definition of the markup μ_{jt} therefore relates it with Hicks-neutral productivity ω_{jt} as

$$\ln \mu_{jt} = p_{jt} - \ln v(K_{jt}, M_{jt}, P_{Ljt}, P_{Mjt}) + \omega_{jt}.$$

From the production function in equation (1), we finally have $\phi(z_{jt}, \delta_{jt}) = \ln F(K_{jt}, L_{jt}, M_{jt}) + \ln \mu_{jt} - p_{jt} + \ln v(K_{jt}, M_{jt}, P_{Ljt}, P_{Mjt})$. Assuming that the price of output P_{jt} is included in z_{jt} (as in De Loecker et al. 2016), it follows that

$$\zeta_{jt} = \mu_{jt} - E(\mu_{jt}|z_{jt}) \tag{10}$$

covaries with any true determinant of the markup that has not been controlled for by observables z_{jt}.

Bias in estimated disturbance. With $\delta_{jt} \neq 0$, DLW obtain an estimate of $\widetilde{\varepsilon}_{jt} = \zeta_{jt} + \varepsilon_{jt}$ and substitute it into equation (3) in lieu of ε_{jt}. However, $\widehat{\widetilde{\varepsilon}}_{jt}$ is a biased estimate of ε_{jt}. Even though the regression in the first step of ACF forces the expectation of $\widehat{\widetilde{\varepsilon}}_{jt}$ conditional on z_{jt} to be zero, equation (10) implies that $\widehat{\widetilde{\varepsilon}}_{jt}$ covaries with any true determinant of the markup that has not been controlled for by observables z_{jt}. The markup obtained by substituting $\widehat{\widetilde{\varepsilon}}_{jt}$ into equation (3) is therefore inversely related with any true determinant of the markup that has not been controlled for by observables z_{jt}.

The economist in the opening paragraph in Section 1 who uses the DLW method and then regresses the estimated markup on a dummy that is one for

the acquiring firms in the "after" period and zero otherwise is a case in point: the bias in $\widehat{\widetilde{\varepsilon}}_{jt}$ predisposes her to finding that the markup of the acquiring firms has decreased rather than increased following the wave of acquisitions.

Bias in estimated output elasticity. With $\delta_{jt} \neq 0$, $\phi(z_{jt})$ in equation (5) becomes $\phi(z_{jt}, \delta_{jt})$. DLW obtain an estimate of $\widetilde{\phi}(z_{jt})$ and substitute it into equation (5) in lieu of $\phi(z_{jt}, \delta_{jt})$. Rewriting equation (5) accordingly yields

$$\begin{aligned} q_{jt} &= \ln F(K_{jt}, L_{jt}, M_{jt}) \\ &\quad + g\left(\widetilde{\phi}(z_{jt-1}) + \zeta_{jt-1} - \ln F(K_{jt-1}, L_{jt-1}, M_{jt-1})\right) + \xi_{jt} + \varepsilon_{jt}. \end{aligned} \tag{11}$$

To take the most favorable case, let ω_{jt} follow an $AR(1)$ process with parameter ρ so that $g(\omega_{jt-1}) = \rho\omega_{jt-1}$ and $\rho\zeta_{jt-1}$ becomes part of the composite error term. Even in this case, however, $\delta_{jt} \neq 0$ invalidates commonly used instruments in the second step of ACF.[10]

Current capital K_{jt} is routinely used as an instrument in the second step of ACF. The underlying assumption is that the firm decides on investment, and thus capital, in period $t-1$ before it observes ω_{jt}; hence, K_{jt} is uncorrelated with ξ_{jt}. Under an analogous assumption, current labor L_{jt} is also often used as an instrument (De Loecker et al. 2016, p. 471). However, none of these instruments is valid with $\delta_{jt} \neq 0$: any variable that is chosen in period $t-1$ is chosen with knowledge of Hicks-neutral productivity ω_{jt-1} and the demand shock δ_{jt-1} and, from equation (9), is therefore correlated with what remains of ω_{jt-1} and δ_{jt-1} in ζ_{jt-1} after controlling for z_{jt-1}.[11] Using invalid instruments biases the estimated output elasticity.

Sign and size of bias in estimated output elasticity. We quantify the sign and size of the bias in the example of a single-input Cobb-Douglas production function $\ln F(K_{jt}, L_{jt}, M_{jt}) = \beta_L l_{jt}$ and an $AR(1)$ process with parameter ρ for Hicks-neutral productivity ω_{jt}. Following ACF and De Loecker

[10] Alternatively, take a Taylor series expansion of the conditional expectation function $g(\cdot)$ around $\zeta_{jt-1} = 0$ and absorb all terms involving powers of ζ_{jt-1} into the composite error term. It is unlikely that the resulting difficulties can be resolved by IV methods (Hausman, Newey & Powell 1995).

[11] Lagged inputs K_{jt-1}, L_{jt-1}, and M_{jt-1} remain valid instruments because they are by construction uncorrelated with ζ_{jt-1} if they are included in z_{jt-1}.

et al. (2016), we assume that current labor L_{jt} is chosen in period $t-1$ and use it as an instrument in the second step of ACF. Finally, the demand the firm faces is simply $\ln D(z_{jt}^D, \delta_{jt}) = -\eta p_{jt} + \delta_{jt}$, where $z_{jt}^D = p_{jt}$, and the vector of observables is $z_{jt} = (l_{jt}, p_{jt})$.

For simplicity, take ρ to be known and write equation (11) as $q_{jt} - \rho\widetilde{\phi}(z_{jt-1}) = \beta_L (l_{jt} - \rho l_{jt-1}) + \rho\zeta_{jt-1} + \xi_{jt} + \varepsilon_{jt}$. In terms of population moments, the IV estimator for the output elasticity β_L is

$$\widehat{\beta}_L^{IV} = \frac{E\left(l_{jt}\left(q_{jt} - \rho\widetilde{\phi}(z_{jt-1})\right)\right)}{E\left(l_{jt}\left(l_{jt} - \rho l_{jt-1}\right)\right)} = \beta_L + \frac{\rho E\left(l_{jt}\zeta_{jt-1}\right)}{E\left(l_{jt}\left(l_{jt} - \rho l_{jt-1}\right)\right)}. \tag{12}$$

The bias in the estimated output elasticity hinges on the correlation between the instrument l_{jt} and $\zeta_{jt-1} = \omega_{jt-1} - E(\omega_{jt-1}|z_{jt-1}) = \delta_{jt-1} - E(\delta_{jt-1}|z_{jt-1})$ (see equation (9)). As the demand for labor likely rises with Hicks-neutral productivity as well as with the demand shock, we expect a positive bias.

In Appendix A, we show that if l_{jt} follows an $AR(1)$ process with parameter ρ_L, then the IV estimator can be written as

$$\begin{aligned} \widehat{\beta}_L^{IV} &= \beta_L \left(1 + \frac{\rho}{1 - \rho\rho_L}\sqrt{\frac{1-R^2}{Var(\beta_L l_{jt})/Var(q_{jt})}}Corr(l_{jt}, \zeta_{jt-1})\right) \\ &= \beta_L\left(1 + bias\right), \end{aligned} \tag{13}$$

where $R^2 = 1 - \frac{Var(\zeta_{jt})}{Var(q_{jt}^*)}$ is the coefficient of determination in the (infeasible) regression of q_{jt}^* on observables z_{jt}.[12] Equation (13) indicates that even if the R^2 is close to 1 in the first step of ACF, the key to the quantitative importance of the bias is the correlation between the instrument l_{jt} and ζ_{jt-1}.[13]

Table 1 takes $\rho = 0.9$, $\rho_L = 0.6$, and $Var(\beta_L l_{jt})/Var(q_{jt}) = 1$ and considers three degrees of correlation between the instrument l_{jt} and ζ_{jt-1} and three values for R^2. The bias in the estimated output elasticity can go from a modest but significant 6 percentage points to an overwhelming 37 percentage points.

Outside our example, the bias in the estimated output elasticity can be large as well. Following DLW, Brandt et al. (2017) estimate the output

[12] Equation (13) assumes a stationary environment and, without loss of generality, that the instrument is in differences with respect to its mean.

[13] The R^2 in the regression of q_{jt} on z_{jt} in the first step of ACF is a lower bound on the R^2 in the regression of q_{jt}^* on z_{jt}.

		R^2		
		0.99	0.95	0.90
$Corr(l_{jt}, \zeta_{jt-1})$	0.30	0.059	0.131	0.186
	0.45	0.088	0.197	0.278
	0.60	0.117	0.262	0.371

Table 1: Bias in estimated output elasticity in percentage points.

elasticities of labor and materials to be 0.05 and 0.91. Jaumandreu & Yin (2018) use the same data as Brandt et al. (2017) but attempt to minimize the impact of unobserved demand heterogeneity. They estimate the output elasticities of labor and materials to be 0.29 and 0.61.

Bias in estimated markup. Substituting biased estimates of the disturbance ε_{jt} and the output elasticity $\beta_X(\cdot)$ into equation (3), DLW obtain

$$\begin{aligned} \mu_{jt}^{DLW} &= \frac{\beta_X(K_{jt}, L_{jt}, M_{jt})(1 + bias_{jt})}{S_{Xjt}^R} \exp(-\varepsilon_{jt} - \widetilde{\varepsilon}_{jt} + \varepsilon_{jt}) \\ &= \mu_{jt}(1 + bias_{jt}) \exp(-\zeta_{jt}), \end{aligned}$$

where we index the bias by j and t to accommodate production functions other than the Cobb-Douglas from our example. It follows that

$$\ln \mu_{jt}^{DLW} \approx \ln \mu_{jt} + bias_{jt} - \zeta_{jt}.$$

The bias in the estimated markup μ_{jt}^{DLW} has two components. The first component affects the unconditional expectation of μ_{jt}^{DLW} and hence its level since

$$E(\ln \mu_{jt}^{DLW}) = \ln \mu_{jt} + E(bias_{jt}). \tag{14}$$

The second component of the bias affects the conditional expectation of μ_{jt}^{DLW} and hence how it correlates with variables that the economist may be interested in such as a firm's export status or measures of trade liberalization. To see this, note that for a variable w_{jt} that is orthogonal to ε_{jt} and not perfectly collinear with z_{jt}, we have

$$E(\ln \mu_{jt}^{DLW} | w_{jt}) = E(\ln \mu_{jt} | w_{jt}) + E(bias_{jt} | w_{jt}) - E(\zeta_{jt} | w_{jt}). \tag{15}$$

5 Testing for the effects of unobserved demand heteroegeneity

In this section, we test whether the first step of ACF is correctly specified as $q_{jt} = \phi(z_{jt}) + \varepsilon_{jt}$ (see again equation (4)) or becomes $q_{jt} = \phi(z_{jt}, \delta_{jt}) + \varepsilon_{jt}$ (see equation (8)). The difficulty is that δ_{jt} is inherently unobservable, as is its correlation with the observables z_{jt}. We overcome this difficulty by exploiting that our data contains a firm- and year-specific indicator of the state of demand (slump, stability, and expansion). This market dynamism variable mdy_{jt} is as good a proxy for shifts in the demand a firm faces as one can hope for in production data and therefore an important component of δ_{jt}. At the same time, there is no reason to believe that it captures differences in demand across firms or time in their entirety. Hence, while our market dynamism variable mdy_{jt} is useful for testing purposes, adding it to the observables z_{jt} is not a solution to the problem of unobserved demand heterogeneity.

Data. Our data come from the Encuesta Sobre Estrategias Empresariales (ESEE) survey, a firm-level survey of the Spanish manufacturing sector, and spans 1990-2012. Appendix B provides details on the sample and variables. We estimate the production function separately for 10 industries.

Specification and estimation. We specify a Cobb-Douglas production function $\ln F(K_{jt}, L_{jt}, M_{jt}) = \beta_0 + \beta_t + \beta_K k_{jt} + \beta_L l_{jt} + \beta_M m_{jt}$, where β_0 is a constant and β_t is a set of 21 year dummies. As in ACF, we specify an $AR(1)$ process with parameter ρ for Hicks-neutral productivity ω_{jt}.

We invert the demand for a variable input and write $\omega_{jt} = h(z_{jt})$. We include input quantities k_{jt}, l_{jt}, and m_{jt}, the real price of labor $p_{Ljt} - p_{jt}$, and the real price of materials $p_{Mjt} - p_{jt}$ in the observables z_{jt} in addition to the constant and the year dummies.

In the first step of ACF, we flexibly approximate $\phi(z_{jt})$ in equation (4) by a complete polynomial of order 3 in the continuous variables included in z_{jt}, the constant, and the year dummies and estimate by OLS. In the second step of ACF, we estimate equation (5) by GMM. The instruments are k_{jt}, k_{jt-1}, l_{jt-1}, m_{jt-1}, and $\widehat{\phi}(z_{jt-1})$ in addition to the constant and the year dummies. We correct the standard errors for the two-step nature of the estimation (see Appendix C).

Markup. While our approach extends directly to alternative assumptions, we assume that both labor L_{jt} and materials M_{jt} are variable inputs.[14] Combining equation (3) for labor and materials yields

$$\mu_{jt} = \frac{\nu(K_{jt}, L_{jt}, M_{jt})}{S^R_{Ljt} + S^R_{Mjt}} \exp(-\varepsilon_{jt}), \tag{16}$$

where $\nu(\cdot) = \beta_L(\cdot) + \beta_M(\cdot) = \frac{\partial \ln F(\cdot)}{\partial l_{jt}} + \frac{\partial \ln F(\cdot)}{\partial m_{jt}}$ is the short-run elasticity of scale.[15] As in DLW, we obtain the markup μ^{DLW}_{jt} of firm j in period t by substituting estimates of the parameters $\nu = \beta_L + \beta_M$ of the Cobb-Douglas production function and of the disturbance ε_{jt} into equation (16).

Results. Table 2 reports the results from the DLW method. Column (1) shows the average (log) markup by industry, along with the sample standard deviation. The average markup ranges from 0.090 in industry 1 to 0.445 in industry 3. Columns (2)–(4) show the underlying production function estimates. The output elasticity of capital β_K is plausible although not significant at the 5% level in industries 5, 6, and 8. The short-run elasticity of scale ν is on the high side and in 7 industries ranges from 0.956 to 1.173.

While the extant literature does not routinely conduct formal specification tests, the Sargan test in column (5) rejects the specification in 3 industries at the 5% significance level. This is not surprising: as shown in Section 4, if there are differences in demand across firms or time that cannot be fully controlled for by z_{jt}, then $\delta_{jt} \neq 0$ and k_{jt} is no longer a valid instrument in the second step of ACF.

Test. To more specifically test for the effects of unobserved demand heterogeneity, recall from equation (11) that if $\delta_{jt} \neq 0$, then equation (5) in the second step of ACF becomes

$$\begin{aligned} q_{jt} &= \ln F(K_{jt}, L_{jt}, M_{jt}) \\ &\quad + \rho\left(\widetilde{\phi}(z_{jt-1}) - \ln F(K_{jt-1}, L_{jt-1}, M_{jt-1})\right) + \rho\zeta_{jt-1} + \xi_{jt} + \varepsilon_{jt}, \end{aligned} \tag{17}$$

[14] See again footnote 3.

[15] Doraszelski & Jaumandreu (2019) and Raval (2020) show that, in practice, the level of the estimated markup and its correlation with variables of interest can be different depending on whether labor or materials is used in equation (3). Combining them in equation (16) ameliorates this problem.

Table 2: DLW method

	Baseline specification				Sargan test		Incl. *mdy*	Regression on *mdy*	
	Markup A (s. dev.)	β_K (s. e.)	ν (s. e.)	ρ (s. e.)	p-val. (1 d.f.)	p-val. (2 d.f.)	Markup B (s. dev.)	Markup A (s. e.)	Markup B (s. e.)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. Metals and metal products	0.090	0.087	0.886	0.792	0.214	0.012	0.104	-0.017*	0.036*
	(0.091)	(0.019)	(0.025)	(0.033)			(0.095)	(0.006)	(0.007)
2. Non-metallic minerals	0.195	0.066	0.964	0.505	0.030	0.013	0.202	-0.003	0.067*
	(0.119)	(0.023)	(0.030)	(0.023)			(0.125)	(0.012)	(0.015)
3. Chemical products	0.445	0.061	1.173	0.953	0.225	0.044	0.246	0.015	0.060*
	(0.166)	(0.029)	(0.117)	(0.102)			(0.169)	(0.017)	(0.018)
4. Agric. and ind. machinery	0.347	0.060	1.102	0.928	0.226	0.005	0.198	-0.008	0.024*
	(0.108)	(0.028)	(0.073)	(0.106)			(0.109)	(0.011)	(0.011)
5. Electrical goods	0.364	0.021	1.089	0.918	0.655	0.023	0.251	-0.003	0.035*
	(0.156)	(0.025)	(0.109)	(0.165)			(0.157)	(0.017)	(0.018)
6. Transport equipment	0.118	0.041	0.917	0.824	0.815	0.248	0.145	0.014	0.061*
	(0.130)	(0.031)	(0.051)	(0.060)			(0.133)	(0.014)	(0.015)
7. Food, drink and tobacco	0.158	0.080	0.905	0.829	0.001	0.000	0.157	-0.067*	-0.040*
	(0.171)	(0.021)	(0.027)	(0.028)			(0.172)	(0.014)	(0.014)
8. Textile, leather and shoes	0.142	0.035	0.965	0.840	0.548	0.000	0.185	0.001	0.049*
	(0.092)	(0.021)	(0.027)	(0.027)			(0.096)	(0.007)	(0.008)
9. Timber and furniture	0.169	0.120	0.988	0.922	0.002	0.001	0.213	-0.012	0.045*
	(0.102)	(0.031)	(0.076)	(0.074)			(0.107)	(0.010)	(0.011)
10. Paper and printing products	0.239	0.072	0.956	0.820	0.131	0.003	0.244	-0.032	0.024
	(0.150)	(0.019)	(0.029)	(0.039)			(0.152)	(0.017)	(0.017)

where $\widetilde{\phi}(z_{jt}) = E\left(\phi(z_{jt}, \delta_{jt})|z_{jt}\right)$ and $\zeta_{jt} = \phi(z_{jt}, \delta_{jt}) - \widetilde{\phi}(z_{jt})$ is the prediction error. Treating our market dynamism variable mdy_{jt} as a component of δ_{jt} and noting that $\delta_{jt} \neq 0$ generally implies $\zeta_{jt} \neq 0$, we test for $\delta_{jt} \neq 0$ by examining the correlation of mdy_{jt} with the composite error $\rho\zeta_{jt-1} + \xi_{jt} + \varepsilon_{jt}$. Adding mdy_{jt} to the instruments used in the second step of ACF, the Sargan test in column (6) detects a significant correlation and rejects the specification in 9 industries at the 5% significance level.

While the Sargan test points to unobserved demand heterogeneity, it may be difficult to detect this problem from a routine examination of the average markup or the coefficient of determination in the first step of ACF. Indeed, the R^2 exceeds 0.99 in all industries.

Bias in estimated markup. As shown in Section 4, unobserved demand heterogeneity causes a bias in the estimated disturbance ε_{jt} and a bias in the estimated output elasticity $\beta_X(\cdot)$. Plugging biased estimates into equation (16), in turn, causes a bias in the estimated markup μ_{jt}^{DLW} that has two components. The first component affects the level of the estimated markup and the second component how it correlates with variables of interest (see again equations (14) and (15)). We illustrate both components in turn.

Starting with the level component, we include mdy_{jt} in z_{jt}, re-estimate equations (4) and (5), and re-compute the markup μ_{jt}^{DLW}.[16] Column (7) of Table 2 shows the average (log) markup by industry. Because mdy_{jt} is only a proxy for δ_{jt}, there is no reason to believe that the estimates are entirely free of bias. Nevertheless, including mdy_{jt} in z_{jt} decreases the markup noticeably in industries 3, 4, and 5 compared to the baseline in column (1).

Turning to the correlation component, we regress $\ln \mu_{jt}^{DLW}$ on our market dynamism variable mdy_{jt} and report the estimated coefficient in Table 2.[17] In the baseline with mdy_{jt} excluded from z_{jt}, the estimated markup is not significantly correlated with market dynamism in 8 industries and significantly negatively correlated with market dynamism in 2 industries (column (8)). In contrast, with mdy_{jt} included in z_{jt}, the estimated markup is significantly positively correlated in 8 industries (column (9)). The latter conveys, as expected, that firms enjoy a higher markup if their demands are expanding rather than contracting.

[16] We also include mdy_{jt} as an instrument in the second step of ACF.

[17] We include a constant and a set of 21 year dummies in this and all subsequent regressions of this type.

As shown in Section 4, this reversal happens because with mdy_{jt} excluded from z_{jt}, the entire effect of demand heterogeneity is left in the estimated disturbance. This estimate of $\widetilde{\varepsilon}_{jt} = \zeta_{jt} + \varepsilon_{jt}$, in turn, is substituted into equation (16) in lieu of ε_{jt} to obtain the markup. Including mdy_{jt} in z_{jt} absorbs a part of demand heterogeneity. The resulting change in the estimated disturbance rectifies the correlation of the estimated markup with market dynamism.

It turns out that if the data has been generated by a Cobb-Douglas production function and mdy_{jt} has been included in z_{jt}, then regressing $\ln \mu_{jt}^{DLW}$ on mdy_{jt} consistently estimates the correlation of the markup with market dynamism despite the fact that the estimated disturbance and the estimated output elasticity are biased. This follows directly from equation (15) and the fact that for a Cobb-Douglas production function the short-run elasticity of scale ν is a constant. Because ν is a constant so is $bias_{jt}$, and thus $E(bias_{jt}|mdy_{jt})$ is a constant that is absorbed into the constant of the regression. Moreover, $E(\zeta_{jt}|mdy_{jt}) = 0$ because mdy_{jt} has been included in z_{jt}. Therefore, $E\left(\ln \mu_{jt}^{DLW}|mdy_{jt}\right) = E\left(\ln \mu_{jt}|mdy_{jt}\right) + const.$

Of course, if a Cobb-Douglas production function is not appropriate for the data at hand, then $E(bias_{jt}|mdy_{jt})$ is generally not a constant and thus cannot be absorbed into the constant of the regression. It follows that the regression cannot consistently estimate the correlation of the markup with market dynamism even if mdy_{jt} has been included in z_{jt} to ensure $E(\zeta_{jt}|mdy_{jt}) = 0$.

6 Concluding remarks and paths forward

To answer a variety of empirical and theoretical questions and provide policy advice, economists would like to have a way to estimate the firm-level markup that is easy to compute and does not require modelling demand and making detailed assumptions about firm conduct. In an attempt to fulfill these desiderata, DLW obtain the markup from the firm's cost minimization problem by substituting in estimates of the output elasticity of a variable input and the disturbance that separates actual from planned output. These estimates are obtained using the OP/LP procedure.

As we have shown, the DLW method does not free the researcher from having to think carefully about the specification of demand and assumptions on firm conduct. In the cost minimization problem, the firm's planned output

Q^*_{jt} summarizes the demand the firm faces. Because planned output Q^*_{jt} is unobserved by the econometrician, the scalar unobservable assumption underlying the OP/LP procedure is violated and it is generally not possible to replace Hicks-neutral productivity ω_{jt} by observables z_{jt}. Our paper has highlighted the underappreciated assumption of the DLW method that to consistently estimate markups, it either has to rule out any differences in demand across firms or time or assume that they can be fully controlled for by observables z_{jt}.

The demand that a firm faces is the fundamental determinant of the markup that the firm charges. It, in turn, depends not only on the product characteristics of the firm but, in imperfectly competitive industries, also on the product characteristics of its rivals and their prices or quantities. Hence, to use an OP/LP procedure to estimate the production function and obtain the markup, the DLW method would have to observe and control for all these determinants of the markup. This is unappealing from a conceptual point of view as it essentially requires controlling for the main object that one is interested in estimating. Because typical production data has even less information on demand than our market dynamism variable and rivals are partially or completely unobserved, attempting to control for any differences in demand across firms or time is also difficult in practice.

We have characterized the bias in the estimates produced by the DLW method that results from unobserved demand heterogeneity. The bias permeates the level of the estimated markup and its correlation with variables of interest. We have provided an empirical application to test for the effects of unobserved demand heterogeneity. The resulting bias is most pronounced in the correlation of the estimated markup with our market dynamism variable. Similar correlations of the estimated markup with variables of interest are often the focus of attention in applications of DLW.

A natural question is if there are alternative approaches to estimation that are robust to any differences in demand across firms or time even if they cannot be fully controlled for by observables z_{jt}. There is a narrow path forward within the proxy variable paradigm. As we have shown, if a Cobb-Douglas production function is appropriate for the data at hand and if the researcher is only interested in the correlation of the estimated markup with a variable of interest, then she can proceed simply by including this variable in the first step of ACF. This amounts to purpose-building the markup for the ex-post analysis and does not address the level component of the bias.

To proceed further and estimate the output elasticity in the second step

of ACF, the researcher can drop instruments such as contemporaneous capital and labor that we have shown to be no longer valid if $\delta_{jt} \neq 0$. This relies on assuming an $AR(1)$ process for Hicks-neutral productivity ω_{jt} (see again footnote 10) but not on assuming a Cobb-Douglas production function. This way of proceeding, in theory, avoids both the level and the correlation component of the bias in the estimated markup. How well it works in practice no doubt depends on the data at hand.

A potentially wider path forward is to take a dynamic panel approach to estimation (Arellano & Bond 1991, Arellano & Bover 1995, Blundell & Bond 1998, Blundell & Bond 2000). The dynamic panel approach is well-established for estimating production functions. Its main advantage is that it avoids the inversion in the first step of ACF and therefore introducing δ_{jt} into the estimation. Hence, the dynamic panel approach is robust to $\delta_{jt} \neq 0$ and offers a solution to the problem of unobserved demand heterogeneity. Moreover, it offers a solution that does not require the researcher to control for any differences in demand across firms or time.

Ackerberg (2020) provides a detailed comparison of the OP/LP procedure and the dynamic panel approach. We highlight two disadvantages of the dynamic panel approach. First, it again relies on the assumption that ω_{jt} follows an $AR(1)$ process. Second, because the dynamic panel approach avoids the inversion in the first step of ACF, it does not yield an estimate of the disturbance ε_{jt}. It is thus not able to estimate the markup μ_{jt} separately from the disturbance ε_{jt} and instead delivers an estimate of $\ln \mu_{jt} + \varepsilon_{jt}$ (obtained by bringing $\exp(-\varepsilon_{jt})$ to the left-hand side in equation (3) or (16)). Any average across groups of firms and/or years involving a sufficiently large number of observations is therefore a consistent estimate of the average (log) markup for these firms and/or years. Despite these disadvantages, there is little reason not to use the dynamic panel approach at least to alleviate concerns about unobserved demand heterogeneity.

A final consideration is that consistently estimating the output elasticity may be difficult in a model that restricts productivity to be single-dimensional. A number of recent papers provide evidence of labor-augmenting productivity (Doraszelski & Jaumandreu 2018, Raval 2019, Zhang 2019, Demirer 2020). In contrast to Hicks-neutral productivity, labor-augmenting productivity directly enters the output elasticity. The literature has only recently begun to develop more sophisticated models and estimators to handle multi-dimensional productivity. Doraszelski & Jaumandreu (2019), Demirer (2020), and Raval (2020) in particular highlight the implications of biased

technological change for markup estimation.

Appendix A

We assume a stationary environment and, without loss of generality, that $E(l_{jt}) = 0$. Moreover, $E(\zeta_{jt}) = 0$ because ζ_{jt} is mean independent of z_{jt}. Letting $E(l_{jt}\zeta_{jt-1}) = Cov(l_{jt}, \zeta_{jt-1})$ and $E(l_{jt}(l_{jt} - \rho l_{jt-1})) = Cov(l_{jt}, l_{jt} - \rho l_{jt-1})$, we write the IV estimator in equation (12) equivalently as

$$\begin{aligned}\widehat{\beta}_L^{IV} &= \beta_L \left(1 + \frac{\rho}{\beta_L} \frac{Cov(l_{jt}, \zeta_{jt-1})}{Cov(l_{jt}, l_{jt} - \rho l_{jt-1})}\right) \\ &= \beta_L \left(1 + \frac{\rho}{\beta_L} \sqrt{\frac{Var(\zeta_{jt})}{Var(l_{jt} - \rho l_{jt-1})}} \frac{Corr(l_{jt}, \zeta_{jt-1})}{Corr(l_{jt}, l_{jt} - \rho l_{jt-1})}\right),\end{aligned}$$

where $Var(\zeta_{jt}) = Var(\zeta_{jt-1})$ because of stationarity.

Note that we can write

$$\frac{1}{\beta_L} \sqrt{\frac{Var(\zeta_{jt})}{Var(l_{jt} - \rho l_{jt-1})}} = \sqrt{\frac{Var(\zeta_{jt})/Var(q_{jt}^*)}{Var(\beta_L l_{jt})/Var(q_{jt}^*)}} \sqrt{\frac{Var(l_{jt})}{Var(l_{jt} - \rho l_{jt-1})}},$$

where $Var(\zeta_{jt})/Var(q_{jt}^*)$ is the proportion of variance of q_{jt}^* explained by ζ_{jt}. This proportion can also be written as $1 - R^2$, where R^2 is the coefficient of determination in the (infeasible) regression of q_{jt}^* on observables z_{jt}. Finally, if l_{jt} follows an $AR(1)$ process with parameter ρ_L, then $Corr(l_{jt}, l_{jt} - \rho l_{jt-1}) = (1 - \rho\rho_L)\sqrt{\frac{Var(l_{jt})}{Var(l_{jt} - \rho l_{jt-1})}}$. Combining expressions yields equation (13).

Appendix B

The ESEE is a firm-level survey of the Spanish manufacturing sector sponsored by the Ministry of Industry. At the beginning of the survey, about 5% of firms with up to 200 workers were sampled randomly by industry and size strata. All firms with more than 200 workers were included in the survey and 70% of these larger firms responded. Firms disappear over time from the sample due to either exit (shutdown or abandonment of activity) or attrition. To preserve representativeness, samples of newly created firms were added to the initial sample almost every year and some additions counterbalanced attrition.

We observe firms for a maximum of 23 years between 1990 and 2012. We restrict the sample to firms with at least three years of observations, giving

a total of 3026 firms and 26977 observations. The number of firms with 3, 4,..., 23 years of data is 398, 298, 279, 278, 290, 324, 122, 111, 137, 96, 110, 66, 66, 98, 66, 40, 37, 44, 37, 42 and 87 respectively.[18]

In what follows we list the variables that we use, beginning with the variables that we take directly from the data source.

- *Revenue* (R). Value of produced goods and services computed as sales plus the variation of inventories.

- *Investment* (I). Value of current investments in equipment goods (excluding buildings, land, and financial assets) deflated by a price index of investment. The price of investment is the equipment goods component of the index of industry prices computed and published by the Spanish Ministry of Industry.

- *Capital* (K). Capital at current replacement values is computed recursively from an initial estimate and the data on investments I at $t-1$ using industry-specific depreciation rates. Capital in real terms is obtained by deflating capital at current replacement values by the price index of investment.

- *Labor* (L). Total hours worked computed as the number of workers times the average hours per worker, where the latter is computed as normal hours plus average overtime minus average working time lost at the workplace.

- *Intermediate consumption* (MB). Value of intermediate consumption or materials' bill.

- *Proportion of white collar workers* (pwc). Fraction of non-production workers.

- *Advertising* (adv). Firm expenditure in advertising.

- *R&D Expenditures* ($R\&D$). Cost of intramural R&D activities, payments for outside R&D contracts with laboratories and research centers, and payments for imported technology in the form of patent li-

[18]Table D1 in Doraszelski & Jaumandreu (2019) shows the industry labels along with their definitions in terms of the ESEE, ISIC and NACE classifications and the number of firms and observations per industry.

censing or technical assistance, with the various expenditures defined according to the OECD Frascati and Oslo manuals.

- *Price of output* (P). Firm-level price index for output. Firms are asked about the price changes they made during the year in up to five separate markets in which they operate. The price index is computed as a Paasche-type index of the responses.

- *Price of labor* (P_L). Hourly wage cost computed as wage bill divided by total hours worked.

- *Price of materials* (P_M). Firm-specific price index for intermediate consumption. Firms are asked about the price changes that occurred during the year for raw materials, components, energy, and services. The price index is computed as a Paasche-type index of the responses.

- *Market dynamism* (mdy). Firms are asked to assess the current and future situation of the main market in which they operate. The demand shifter codes the responses as 0, 0.5, and 1 for slump, stability, and expansion, respectively.

We construct a number of additional variables. We consistently subtract advertising from intermediate consumption because it is not a production input. We define variable cost as the wage bill plus the cost of intermediate consumption (minus advertising), minus the R&D expenditures and an estimate of the part of the wage bill corresponding to white collar workers. The estimation assumes that white-collar employees work the same number of hours but have an average wage 1.25 times higher. This is important to better approximate variable cost.

- *Output* (Q). Revenue deflated by the firm-specific price index of output.

- *Materials* (M). Value of intermediate consumption minus advertising deflated by the firm-specific price index of materials.

- *Variable cost* (VC). Wage bill (including social security payments) plus the cost of intermediate consumption minus advertising, R&D expenditures, and white collar pay.

Appendix C

Let γ be the parameters estimated in equation (4) in the first step of ACF and θ the parameters estimated in equation (5) in the second step of ACF. Given the estimate $\widehat{\gamma}$, $\xi_{jt} + \varepsilon_{jt} = r_{jt}(\theta, \phi(z_{jt-1}; \widehat{\gamma}))$ in equation (5) depends on $\widehat{\phi}(z_{jt-1}) = \phi(z_{jt-1}; \widehat{\gamma})$. Stacking yields the $T_j \times 1$ vector $r_j(\theta, \phi(z_{j,-1}; \widehat{\gamma}))$, where T_j is the number of observations for firm j.

Following Wooldridge (2010), let

$$D_0 = E[w_j' r_j(\theta_0, \phi(z_{j,-1}; \widehat{\gamma})) r_j(\theta_0, \phi(z_{j,-1}; \widehat{\gamma}))' w_j]$$

be the variance of the orthogonality conditions based on the $T_j \times Q$ matrix of instruments w_j in the second step of ACF, evaluated at the true value of θ. Expanding $r_j(\cdot)$ around the true value of γ yields $r_j(\theta_0, \phi(z_{j,-1}; \widehat{\gamma})) \approx r_j(\theta_0, \phi(z_{j,-1}; \gamma_0)) + \frac{\partial r_j}{\partial \phi} \nabla_\gamma \phi(z_{j,-1}; \gamma_0)(\widehat{\gamma} - \gamma_0)$. Since γ is estimated by OLS, we use $(\widehat{\gamma} - \gamma_0) = \sum_j (f(z_j)' f(z_j))^{-1} f(z_j)' \varepsilon_j$, where $f(z_j)$ are the regressors in the first step of ACF, and replace $r_j(\cdot)$ by

$$\widetilde{r}_j(\theta_0, \gamma_0, \varepsilon_j) = r_j(\theta_0, \phi(z_{j,-1}; \gamma_0)) + \frac{\partial r_j}{\partial \phi} \nabla_\gamma \phi(z_{j,-1}; \gamma_0) \sum_j (f(z_j)' f(z_j))^{-1} f(z_j)' \varepsilon_j.$$

Replacing the true values of θ, γ, and ε_j by their estimates, we estimate D_0 as $\widehat{D} = \frac{1}{N} \sum_j w_j' \widetilde{r}_j(\widehat{\theta}, \widehat{\gamma}, \widehat{\varepsilon}_j) \widetilde{r}_j(\widehat{\theta}, \widehat{\gamma}, \widehat{\varepsilon}_j)' w_j$. Next, we use $\widehat{D}$ in the usual sandwich formula for the asymptotic variance of the estimated parameters θ and in the optimal weighting matrix to compute the Sargan test.

References

Ackerberg, D. (2020), Timing assumptions and efficiency: Empirical evidence in a production function context, Working paper, University of Texas, Austin.

Ackerberg, D., Benkard, L., Berry, S. & Pakes, A. (2007), Econometric tools for analyzing market outcomes, *in* J. Heckman & E. Leamer, eds, 'Handbook of Econometrics', Vol. 6A, North-Holland, Amsterdam, pp. 4171–4276.

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

Arellano, M. & Bond, S. (1991), 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies* **58**, 277–297.

Arellano, M. & Bover, O. (1995), 'Another look at instrumental variable estimation of error component models', *Journal of Econometrics* **68**, 29–52.

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

Bain, J. (1951), 'Relation of profit rate to industry concentration: American manufacturing, 1936-1940', *Quarterly Journal of Economics* **65**, 293–324.

Baker, J. & Bresnahan, T. (1985), 'The gains from merger or collusion in product-differentiated industries', *Journal of Industrial Economics* **33**, 59–76.

Berry, S., Gaynor, M. & Scott-Morton, F. (2020), 'Do increasing markups matter? Lessons from empirical industrial organization', *Journal of Economic Perspectives* **33**(3), 44–68.

Berry, S., Levinsohn, J. & Pakes, A. (1995), 'Automobile prices in market equilibrium', *Econometrica* **63**(4), 841–890.

Blundell, R. & Bond, S. (1998), 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of Econometrics* **87**, 115–143.

Blundell, R. & Bond, S. (2000), 'GMM estimation with persistent panel data: An application to production functions', *Econometric Reviews* **19**, 321–340.

Bond, S., Hashemi, G., Kaplan, G. & Zoch, P. (2020), Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data, Working paper no. 27002, NBER, Cambridge.

Brandt, L., Van Biesebroeck, J., Wang, L. & Zhang, Y. (2017), 'WTO accession and performance of Chinese manufacturing firms', *American Economic Review* **107**(9), 2784–2820.

Brandt, L., Van Biesebroeck, J., Wang, L. & Zhang, Y. (2019), 'WTO accession and performance of Chinese manufacturing firms: Corrigendum', *American Economic Review* **109**(4), 1616–1621.

De Loecker, J. & Eeckhout, J. (2018), Global market power, Working paper no. 24768, NBER, Cambridge.

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

De Loecker, J., Goldberg, P., Khandelwal, A. & Pavcnik, N. (2016), 'Prices, markups and trade reform', *Econometrica* **84**(2), 445–510.

De Loecker, J. & Scott, P. (2016), Estimating market power: Evidence from the US brewing industry, Working paper no. 22957, NBER, Cambridge.

De Loecker, J. & Warzynski, F. (2012), 'Markups and firm-level export status', *American Economic Review* **102**(6), 2437–2471.

Demirer, M. (2020), Production function estimation with factor-augmenting technology: An application to markups, Working paper, MIT, Cambridge.

Doraszelski, U. & Jaumandreu, J. (2018), 'Measuring the bias of technological change', *Journal of Political Economy* **126**(3), 1027–1084.

Doraszelski, U. & Jaumandreu, J. (2019), Using cost minimization to estimate markups, Working paper no. 14114, CEPR, London.

Foster, L., Haltiwanger, J. & Syverson, C. (2008), 'Reallocation, firm turnover, and efficiency: Selection on productivity or profitability?', *American Economic Review* **98**(1), 394–425.

Hall, R. (1988), 'The relation between price and marginal cost in U.S. industry', *Journal of Political Economy* **96**(5), 921–947.

Hausman, J., Newey, W. & Powell, J. (1995), 'Nonlinear errors in variables. estimation of some Engel curves', *Journal of Econometrics* **65**, 205–213.

Jaumandreu, J. & Yin, H. (2018), Cost and product advantages: Evidence from Chinese manufacturing firms, Working paper no. 11862, CEPR, London.

Levinsohn, J. & Petrin, A. (2003), 'Estimating production functions using inputs to control for unobservables', *Review of Economic Studies* **70**(2), 317–341.

Olley, S. & Pakes, A. (1996), 'The dynamics of productivity in the telecommunications industry', *Econometrica* **64**(6), 1263–1297.

Pakes, A. (1994), Dynamic structural models, problems, and prospects: mixed continuous discrete controls and market interactions, *in* C. Sims, ed., 'Advances in econometrics: Sixth World Congress', Vol. 2, Cambridge University Press, Cambridge.

Raval, D. (2019), 'The micro elasticity of substitution and non-neutral technology', *Rand Journal of Economics* **50**(1), 147–167.

Raval, D. (2020), Testing the production function approach to markup estimation, Working paper, Federal Trade Commission, Washington.

Samuelson, P. (1947), *Foundations of economic analysis*, Harvard University Press, Cambridge.

Wooldridge, J. (2010), *Econometric analysis of cross section and panel data*, 2nd edn, MIT Press, Cambridge.

Zhang, H. (2019), 'Non-neutral technology, firm heterogeneity, and labor demand', *Journal of Development Economics* **140**, 145–168.