Dangerous Shortcuts: Ignoring Marginal Cost Determinants in Markup Estimation*

Jordi Jaumandreu†

Boston University and CEPR

This version: June 2018

First draft: June 2017

Preliminary and incomplete

Abstract

De Loecker and Warzynski (2012) proposed a way to estimate markups under imperfect competition from firm-level data. The basic idea is first to estimate the elasticity of output with respect to a variable factor of production, and then divide this elasticity by the share of the input expenditure in the relevant sales. This note argues that this way to estimate markups is affected by a problem of circularity: under imperfect competition, to estimate the elasticity of an input you need to know, or estimate at the same time, the markups. The application of the proposal, ignoring this fact, is likely to produce inconsistent estimates and generate large biases in inferences about markups.

*I thank Uli Dosraszelski, Ivan Fernandez-Val and Shuheng Lin for useful discussions.

†Department of Economics, 270 Bay State Road, MA02215, USA. Email: jordij@bu.edu.

1. Introduction

De Loecker and Warzynski (2012), henceforth DLW, proposed a way to estimate markups under imperfect competition from firm-level data. The basic idea is first to estimate the elasticity of output with respect to a variable factor of production, and then divide this elasticity by the share of the input expenditure in the relevant sales. The procedure has been recently applied by De Loecker, Goldberg, Khandelval and Pavcnick (2016), and is becoming increasingly popular. This note argues that this way to estimate markups is affected by a problem of circularity: under imperfect competition, to estimate the elasticity of an input you need to know, or estimate at the same time, the markups.[1] The application of the proposal, ignoring this fact, is likely to produce inconsistent estimates and generate large biases in inferences about markups.[2]

Let's anticipate intuitively why this happens. Having prices, markups follow if one can estimate marginal cost.[3] The idea is that the output elasticity of a variable input can be used to infer production marginal cost. The problem is that, in the standard setting and from cost minimization, the elasticity of no input can be estimated consistently if marginal cost is not observed or, at least, simultaneously estimated.[4] In particular, when unobservable marginal

[1] The reader at this point may ask with surprise how is possible that this has gone unnoticed until now. The answer probably lies in the loose treatment of output and input prices in the literature on structural estimation of production functions. This loose treatment has been favored in turn by the scarcity of price data.

[2] Both mentioned articles reach unconventional conclusions. DLW finds that the markups of Eslovenian exporters are, for the same products in a more competitive market, greater than the markups of the non-exporters. De Loecker, Goldberg, Khandelval and Pavcnick (2016) find that Indian firms increased their market power after a tariff liberalization. The lack of reliability of the markup measurements raises doubts on the robustness of the results.

[3] DLW suggest that their way to estimate markups gives also valid inferences on the differences of these markups in the absence of firm-level prices (although they recognize possible problems when varying input elasticities are used). What follows will demonstrate that the absence of firm-level prices induces important additional biases. Moreover, the negative correlation argued by Klette and Griliches (1996) between firm prices and inputs only holds under their assuption that demand heterogeneity is iid.

[4] This doesn't happen under perfect competition because marginal cost is observed through the price.

cost is replaced by its determinants one cannot ignore demand heterogeneity without the risk of biases.[5] The problem becomes more acute if there are adjustment costs of any variable input (e.g. labor) because they constitute a non-ignorable part of the input prices and the own marginal cost.

In Section 2 we restate the setting for the estimation of markups. Section 3 explains the need to deal with marginal cost. Section 4 briefly discusses the added difficulties when there are adjustment costs. Section 5 suscintly compares with profit maximization, that can be also dynamic. Section 6 analyzes the generated biases. Section 7 sketches the alternative suggested in Jaumandreu and Yin (2016) for the case of monopolistic competition, and Jaumandreu and Lin (2016) in a more general case. Appendix A develops an example to sign the bias.

2. The setting for estimating markups

First, let's make clear what is understood by the standard setting. Assume, as in DLW, that firms compete in an imperfectly competitive market in an unspecified way.[6] Only the assumption that firms minimize costs is going to be used. Firm j produces with a given amount of capital, and freely variable amounts of labor and materials, with production function $q_{jt}^P = \ln F(K_{jt}, L_{jt}, M_{jt}) + \omega_{jt}$. Variable ω_{jt} represents Hicks-neutral productivity, that evolves over time as a first order Markov process, $\omega_{jt} = g(\omega_{jt-1}) + \xi_{jt}$, and is observed by the firm but unobserved by the econometrician.[7] Observed production is $q_{jt} = q_{jt}^P + \varepsilon_{jt}$. The disturbance ε_{jt}, uncorrelated with the inputs, accounts for the difference between output q^P and observed output. Variable inputs have been chosen to produce the "planned" quantity q^P.[8] Capital is taken as given to account for the fact that its adjustment is particularly

[5] See Jaumandreu and Yin (2018) for an account on the existence and importance of demand heterogeneity.

[6] Products, in particular, can be homogeneous or differentiated.

[7] The Markov process can be exogenous or endogenous, the latter including some variable that shifts the expected value of productivity as expenditures in R&D, exports or innovation dummies. See, for example, Doraszelski and Jaumandreu (2013), De Loecker (2013) and Peters, Roberts, Vuong and Fryges (2017).

[8] The most natural way of interpreting the difference between q_{jt} and q_{jt}^P is as picking up the unpredictable shocks at the time of setting the quantities of the variable inputs. It can also be interpreted as a pure

costly and requires time.[9] At some point of the discussion we will consider adjustment cost of the input labor while we keep materials completely variable.[10]

The work by Hall (1990), drawing on Solow (1957), showed that, under cost minimization, the FOC of a variable input (materials, say),

$$MC_{jt}\frac{\partial Q_{jt}}{\partial M_{jt}} = P_{Mjt},$$

can be written as

$$MC_{jt} = \frac{\frac{P_{Mjt}M_{jt}}{Q_{jt}^P}}{\beta_{Mjt}},$$

where $\beta_{M_{jt}} \equiv \frac{M_{jt}}{Q_{jt}^P}\frac{\partial Q_{jt}}{\partial M_{jt}}$ stands for the elasticity of output with respect to materials. If we observe the materials bill and the quantity of output that the materials were set to produce, having the elasticity of the input amounts to have marginal cost and conversely. The FOC can be also used to produce directly markups $\mu_{jt} = P_{jt}/MC_{jt}$ by writing

$$\mu_{jt} = \frac{\beta_{Mjt}}{S_{Mjt}^P},$$

where $S_M^P = \frac{P_{Mjt}M_{jt}}{P_{jt}Q_{jt}^P}$ is the relevant revenue share.

3. Dealing with non-observable marginal cost

The elasticity β_{Mjt} is, however, non-observed, so it should be estimated through the estimation of the production function. The estimation of the production function will also provide an empirical measurement of Q_{jt}^P. DLW propose estimating β_{Mjt} by means of an Olley and Pakes (1996)/ Levisohn and Petrin (2003) procedure, henceforth OP/LP, implemented in the form proposed by Ackerberg, Caves and Frazer (2015), from now on ACF. An OP/LP procedure consists of replacing the unobservable ω_{jt} in the production

measurement error.

[9] The assumptions on the production function are the general assumptions used in the literature on stuctural estimation of production functions and productivity: see, for example, Olley and Pakes (1996), Levinshon and Petrin (2003), Akerberg, Caves and Frazer (2015) and Doraszelski and Jaumandreu (2013).

[10] Materials are however not likely to be completely variable because of the presence of subcontracting of parts and pieces (see Doraszelski and Jaumandreu, 2018).

function by an inverted input demand. The implementation proposed by ACF estimates in a first step the disturbance ε_{jt} that separates observed output from relevant output, throwing all the arguments of the production function and the inverted demand in a nonparametric estimation of $q^P = \phi(\cdot)$.

DLW proposal overlooks that, under varying markups, no OP/LP method of estimation can produce a consistent estimate of β_{Mjt} and the residual ε_{jt}, regardless whether an ACF or any other implementation form is used, except in the case that markups are estimated simultaneously with the elasticity. This is true either in the case of static cost minimization or in the case of cost minimization with adjustment costs. It becomes transparent with profit maximization and dynamic pricing. As stated before, we start the discussion with static cost minimization.

Our argument proceeds in two stages. First, given cost minimization we demonstrate that either we observe marginal cost (what amounts to observe the markups) or we need to replace Q_{jt}^P by its determinants. Second, we show that the replacement of Q_{jt}^P by its determinants violates the "scalar unobservable assumption" needed for the consistency of OP/LP procedures (Ackerberg, Benkard, Berry and Pakes, 2007).

Equations generated by cost minimization.

The problem of cost minimization is

$$Min_{L_{jt},M_{jt}} \;\; W_{jt}L_{jt} + P_{jt}M_{jt}, \;\; s.t. \;\; F(K_{jt}, L_{jt}, M_{jt}) = Q_{jt}^e / \exp(\omega_{jt}).$$

The first order conditions are

$$\begin{aligned} MC_{jt}\frac{\partial F(K_{jt}, L_{jt}, M_{jt})}{\partial L_{jt}} \exp(\omega_{jt}) &= W_{jt}, \\ MC_{jt}\frac{\partial F(K_{jt}, L_{jt}, M_{jt})}{\partial M_{jt}} \exp(\omega_{jt}) &= P_{Mjt}, \end{aligned}$$

where we write the Lagrange multiplier of the problem as marginal cost MC_{jt} . Let's briefly derive, for the sake of the discussion, the demand for the inputs conditional on output, variable cost (the function value) and marginal cost (the derivative of the function value). Using the ratio of first order conditions together with the production function we can solve for the output-conditional demand for the variable inputs

$$M_{jt} = M(K_{jt}, \frac{W_{jt}}{P_{Mjt}}, \frac{Q_{jt}^P}{\exp(\omega_{jt})}),$$
$$L_{jt} = L(K_{jt}, \frac{W_{jt}}{P_{Mjt}}, \frac{Q_{jt}^P}{\exp(\omega_{jt})}).$$

Plugging these demands into the objective function, variable cost turns out to be

$$VC_{jt} = VC(K_{jt}, W_{jt}, P_{Mjt}, \frac{Q_{jt}^P}{\exp(\omega_{jt})}),$$

and hence marginal cost is

$$MC_{jt} = \frac{\partial VC_{jt}}{\partial Q_{jt}^P} = VC_4(K_{jt}, W_{jt}, P_{Mjt}, \frac{Q_{jt}^P}{\exp(\omega_{jt})}) \exp(-\omega_{jt}).$$

Notice that all these expressions are implications of the primitive assumptions (cost minimization, fixed capital and variable labor and materials, Hicks-neutral productivity). No functional form assumption has been added.

Which among these expressions can be inverted to substitute for ω_{jt} in the production function? Ideally one would like to rely in the first order conditions, where there is no Q_{jt}^P, but then marginal cost is not observable and must be replaced.[11] If one replaces MC_{jt} by its expression, $\exp(\omega_{jt})$ drops from both FOCs leaving only the ratios $\frac{Q_{jt}^P}{\exp(\omega_{jt})}$.[12] The FOCs written in this way have the same problem that the conditional demands, we can invert them for ω_{jt} but the resulting expressions are going to be conditional on Q_{jt}^P.[13] But output

[11] Can be here useful the fact that MC_{jt} is observable up to a set of parameters? For example $MC_{jt} = \frac{AMC_{jt}}{\beta_{Mjt}} \exp(\varepsilon_{jt})$, where $AMC_{jt} = \frac{P_{Mjt}M_{jt}}{Q_{jt}}$ or average cost of materials. Also $MC_{jt} = \frac{AVC_{jt}}{\beta_{Ljt}+\beta_{Mjt}} \exp(\varepsilon_{jt})$ where $AVC_{jt} = \frac{VC_{jt}}{Q_{jt}}$. In both cases the disturbance ε_{jt} violates the "scalar unobservable assumption." However, notice that this provides an estimable model at the price of giving up the nonlinearity of the Markov process.

[12] Can improve the things the assumption that MC_{jt} is independent of output? In this case $VC_{jt} = c(K_{jt}, W_{jt}, P_{Mjt}) \frac{Q_{jt}^P}{\exp(\omega_{jt})}$ and $MC_{jt} = c(K_{jt}, W_{jt}, P_{Mjt}) \exp(-\omega_{jt})$. It is easy to see that ω_{jt} drops from the FOCs, so we can only rely on the conditional input demands.

[13] One may wonder what happens if the ratio $\frac{Q_{jt}^P}{\exp(\omega_{jt})}$ is replaced inverting for this ratio one of the conditional demands, materials say. Interestingly enough, one gets an expression for MC_{jt} with no other unobservables than ω_{jt}, $MC_{jt} = \overline{MC}(K_{jt}, M_{jt}, W_{jt}, P_{Mjt}) \exp(-\omega_{jt})$. But the expression is not useful for our current purposes since ω_{jt} drops from the FOC.

Q^P_{jt} is also non-observable. In fact Q^P_{jt} is just exactly what we want to estimate in the first stage of the ACF implementation.[14] An OP/LP procedure without observing MC_{jt} needs to replace Q^P_{jt} by its determinants.

Using observables to replace marginal cost and planned output.

Replacing Q^P_{jt} by its determinants seems to be in fact the route taken by DLW, although never is explicitly stated. There is a key ambiguity when DLW say "We follow Levinsohn and Petrin (2003) and rely on material demand

$$m_{jt} = m_t(k_{jt}, \omega_{jt}, z_{jt}),$$

to proxy for productivity by inverting $m_t(\cdot)$, where we collect additional variables potentially affecting optimal input demand choice in the vector z_{jt}."[15] According to our discussion the list of variables that has to be included in z_{jt} is very precise: the input prices W_{jt} and P_{Mjt} as well as all the determinants of Q^P_{jt}. Here is exactly where the problem resides, demanded Q^P_{jt} is in general a function $Q^P_{jt} = Q(P_{jt}, z^S_{jt}, \delta_{jt})$,[16] where z^S_{jt} stands for observed demand shifters and δ_{jt} is unobserved demand heterogeneity. Vector z_{jt} of observable variables can then be specified as $z_{jt} = (w_{jt}, p_{M_{jt}}, p_{jt}, z^S_{jt})$ but, in addition, the presence of unobserved demand heterogeneity δ_{jt} has to be accounted for. The demand for materials becomes

$$m_{jt} = m_t(k_{jt}, \omega_{jt}, z_{jt}, \delta_{jt}).$$

[14] In fact notice that we are going to include Q^P_{jt} as an argument of $\phi(\cdot)$, so we will have $Q^P_{jt} = \phi(., Q^P_{jt}) = \phi(., Q(\cdot)) \equiv Q(\cdot)$ and the arguments of $\phi(\cdot)$ will be only the argumants of $Q(\cdot)$.

[15] Levinsohn and Petrin (2003) proposed in fact to use $m_{jt} = m_t(k_{jt}, \omega_{jt})$ noticing that "input and output prices are assumed to be common across firms (they are suppressed)" (page 322). They had in mind a competitive industry, as Appendix A of page 338 makes clear. Under imperfect competition, equal output and input prices is neither necessary nor sufficient condition to be able to write an expression as above. Under profit maximization, for example, if prices are equal we still need the equality of the markups (see below). Ackerberg, Caves and Frazer (2015) recognize the difficulty to integrate imperfectly competitive firms in their framework to estimate production functions and redirect the reader to the approach that we are criticizing (page 2436).

[16] Planned producction can however be different from expected demand if the firm expects to cover part of the demand with inventories or, on the contrary, plans to accumulate some.

Unfortunately, the demand heterogeneity embodied in δ_{jt} violates the "scalar unobservable" assumption that we need for an OP/LP method to work.[17]

Inverting this demand we get

$$\omega_{jt} = h(k_{jt}, m_{jt}, z_{jt}, \delta_{jt}).$$

Substituting this expression for ω_{jt} in the production function, and collecting all observable variables in the vector x_{jt},[18] it becomes clear that the first stage of the ACF procedure has to estimate

$$q_{jt} = \phi(x_{jt}, \delta_{jt}) + \varepsilon_{jt}.$$

As we do not observe δ_{jt}, the regression will produce estimates of the $\phi(\cdot)$ function and the residuals ε_{jt} that differ from the true values. Notice that this is true regardless whether δ_{jt} is correlated or not with the included variables, although in general we should expect to be correlated.

4. Dynamic cost minimization

There can be adjustment costs of an input, and then the problem of cost minimization is dynamic. Assume, without loss of generality, that only labor has adjustment costs. Jaumandreu and Lin (2017) show that the relevant labor FOC for cost minimization can be written as

$$MC_{jt}\frac{\partial Q_{jt}}{\partial L_{jt}} = W_{jt}(1 + \Delta_{jt}),$$

where Δ_{jt} represents the gap between the wage and the shadow price of labor under adjustment costs (in practice it measures the extent to which adjustment costs are not included

[17] Demand heterogeneity is currently recognized as an important fact. The first paper dealing with demand in the context of production function estimation, Klette and Griliches (1996), assumed δ_{jt} an uncorrelated error. This is also the assumption in DeLoecker (2011) once $\mathbf{z}_{jt}^S$ has been accounted for. Today δ_{jt} is recognized as an autocorrelated unobservable, as important as ω_{jt} or even more, and correlated with it. See Jaumandreu and Yin (2018) for a review of the relevant literature and an empirical assessment.

[18] Notice that, as previously remarked, $\phi(\cdot)$ is no longer a function of other observable variables than the arguments of $Q(\cdot)$. In our notation, $x_{jt} = (p_{jt}, z_{jt}^S)$.

in the variation of the observed wage). The relevant price of labor can be now thought of $W_{jt}^{*} = W_{jt}(1+\Delta_{jt})$ and marginal cost includes now the unobservable Δ_{jt}. For example, Jaumandreu and Lin (2017) show that the relationship marginal cost-observed average variable cost becomes

$$MC_{jt} = \frac{1}{\nu_{jt}} AVC_{jt}(1 + s_{Ljt}\Delta_{jt}^{L}),$$

where s_{Ljt} is the share of wages in observable variable costs. It is not only that the computation of the markup cannot be based in the FOC for labor, is that the estimation of any elasticity is going to be afected by inconsistency if the cost of labor is mispecified and this cyclical determinant of marginal cost omitted. The absence of control for this unobservable creates a problem similar to the omision of control for δ_{jt}.

5. Profit maximization

Cost minimization doesn't allow to consistently estimate input elasticities without knowing, or estimating at the same time, markups. Is this possible if we assume profit maximization? The answer is no, provided that markups vary. Proft maximization produces the conditions

$$\begin{aligned} MR_{jt}\frac{\partial F(K_{jt}, L_{jt}, M_{jt})}{\partial L_{jt}} \exp(\omega_{jt}) &= W_{jt}, \\ MR_{jt}\frac{\partial F(K_{jt}, L_{jt}, M_{jt})}{\partial M_{jt}} \exp(\omega_{jt}) &= P_{M_{jt}}. \end{aligned}$$

To replace ω_{jt} we can use any of the FOCs or, provided that returns to scale are nonconstant, we can solve the system and get the unconditional demand for materials, say,

$$M_{jt} = M(K_{jt}, \frac{W_{jt}}{MR_{jt}}, \frac{P_{M_{jt}}}{MR_{jt}}, \omega_{jt}).$$

To implement any of these solutions we must be able to specify MR_{jt}. Since $MR_{jt} = MC_{jt}$ we can always write $MR_{jt} = P_{jt}(1 + \frac{1-\mu_{jt}}{\mu_{jt}})$, what makes clear that having prices available we still need to know the markups. Expression $(1 + \frac{1-\mu_{jt}}{\mu_{jt}})$ can be written without loss of generality as $(1 - \frac{1}{\eta_{jt}})$, where η_{jt} is the "shadow" elasticity of demand or elasticity of demand thet will cause the firm to set the observed markup if it were pricing statically as a monopolist of its product. In any case, there is no hope to consistently estimate input elasticities without controlling for this heterogeneity. Doraszelski and Jaumandreu (2013) faced this problem estimating $\ln(1-\frac{1}{\eta_{jt}})$ as an argument of the demand for the input, modeling η_{jt} as an unknown function of the firm-level ouput price and the intercept of the output demand.

The relevant elasticity may have different interpretations. It can be assumed to coincide with the individual demand elasticity, as it is the case in monopolistic competition. But it may be also a function of the aggregate elasticity of demand and market structure measures if the market is product-homogeneous. In Cournot competition, for example, $\eta_{jt}=\frac{\eta}{s_{jt}}$ where s_{jt} is the market share of the firm. Nothing precludes this shadow elasticity to emerge from a complex price competition oligopoly model or being a complicated function of the state variables of a dynamic pricing problem (see Jaumandreu and Lin, 2017).

The flexible estimation of these elasticities may alleviate partially the problem of specification of marginal cost. It can be, as we show later, a way to avoid the estimation of the heterogeneity of demand. But it cannot avoid the need to be specific, for example, about the cyclical properties of elasticity and the ratio marginal-average cost (closely related to adjustment costs).

6. Econometric biases

Let's take the case of δ_{jt} as the relevant unobservable (in practice it can also be Δ_{jt} or both δ_{jt} and Δ_{jt}). The first stage of the ACF procedure should be based on the regression

$$q_{jt}=\phi(x_{jt},\delta_{jt})+\varepsilon_{jt}.$$

However, because the heterogeneity embedded in δ_{jt} is unobservable, the first stage can only rely on the integrated conditional expectation

$$q_{jt}=\widetilde{\phi}(x_{jt})+r_{jt}+\varepsilon_{jt}=\widetilde{\phi}(x_{jt})+\widetilde{\varepsilon}_{jt},$$

where $\widetilde{\varepsilon}_{jt}$ is by construction mean-independent of x_{jt}. The feasible second stage of ACF will have to be based on the regression

$$q_{jt} = \ln F(K_{jt}, L_{jt}, M_{jt}) + g(\widetilde{\phi}(x_{jt-1}) + r_{jt-1} - \ln F(K_{jt-1}, L_{jt-1}, M_{jt-1})) + \xi_{jt} + \varepsilon_{jt}.$$

Notice that this regression includes a function $\widetilde{\phi}(x_{jt-1})$ different from $\phi(x_{jt}, \delta_{jt})$, and unobserved heterogeneity r_{jt-1} presumably correlated with K_{jt}, L_{jt} and M_{jt}, inside the function $g(\cdot)$. There is no hope to get β_{Mjt} by standard methods. Let's denote $\widetilde{\beta}_{Mjt}$ the elasticity obtained from this regression ignoring the problem of unobserved heterogeneity and let us use it together with $\widetilde{\varepsilon}_{jt}$ to construct the markup

$$\widetilde{\mu}_{jt} = \frac{\widetilde{\beta}_{Mjt}}{S_{Mjt}} \exp(-\widetilde{\varepsilon}_{jt}),$$

or, in logs,

$$\ln \widetilde{\mu}_{jt} = \ln \frac{R_{jt}}{P_{Mjt} M_{jt}} + \ln \widetilde{\beta}_{Mjt} - \widetilde{\varepsilon}_{jt}.$$

It is easy to see that $\ln \frac{R_{jt}}{P_{Mjt} M_{jt}} = \ln \mu_{jt} - \ln \beta_{Mjt} + \varepsilon_{jt}$,[19] so the constructed markup will have the relationship that follows with the true markup

$$\ln \widetilde{\mu}_{jt} = \ln \mu_{jt} + (\ln \widetilde{\beta}_{Mjt} - \ln \beta_{Mjt}) + (\varepsilon_{jt} - \widetilde{\varepsilon}_{jt}) = \ln \mu_{jt} + u_{jt}.$$

The unconditional expectation of the constructed markup is biased since

$$E(\ln \widetilde{\mu}_{jt}) = \ln \mu_{jt} + E(u_{jt}) = \ln \mu_{jt} + E(\ln \widetilde{\beta}_{Mjt} - \ln \beta_{Mjt}),$$

where to obtain the $E(u_{jt})$ we use that $E(\varepsilon_{jt}) = E(\widetilde{\varepsilon}_{jt}) = 0$. In Appendix A we develop an example that suggests a positive bias under the assumption of a positive relationship between the input and heterogeneity of demand δ_{jt}. On the other hand the unobservable Δ_{jt}, component of the price of the labor input, is likely to be negatively related to M_{jt}, what suggests a negative bias.

[19] $\ln \frac{R_{jt}}{P_{Mjt} M_{jt}} = \ln \frac{P_{jt}}{\frac{P_{Mjt} M_{jt}}{Q_{jt}^P} \frac{Q_{jt}^P}{Q_{jt}}} = \ln \frac{P_{jt}}{\beta_{Mjt} MC_{jt} \exp(-\varepsilon_{jt})} = \ln \mu_{jt} - \ln \beta_{Mjt} + \varepsilon_{jt}.$

Call now d_{jt} to some variable for which we want to test whether it is associated to a different markup level and has been included in the vector x_{jt} (as, for example, the export status of DLW). The conditional expectation of the constructed markup is biased since

$$E(\ln \widetilde{\mu}_{jt}|d_{jt}) = E(\ln \mu_{jt}|d_{jt}) + E(u_{jt}|d_{jt}),$$

where $E(u_{jt}|d_{jt}) = E(\ln \widetilde{\beta}_{Mjt} - \ln \beta_{Mjt}|d_{jt}) + E(\varepsilon_{jt}|d_{jt}) \neq 0$. We drop the term $E(\widetilde{\varepsilon}_{jt}|d_{jt}) = 0$.

Notice that first term of the bias is independent of the type of variation of $\widetilde{\beta}_{Mjt}$ and β_{Mjt}. The estimated parameter may be a constant, i. e. $\widetilde{\beta}_{Mjt} = \widetilde{\beta}_M$, and we can still have a varying β_{Mjt}. The second term of the bias shows that the possible endogeneity of variable d_{jt} in the explanation of Q_{jt} cannot be ignored because the employed procedure has failed to control for ε_{jt}.

In summary, both the level of the markups and its variation according to key variables of interest are likely to be biased.

7. Sketch of solution

As suggested in Jaumandreu and Yin (2018) for the monopolistic competition case, and in Jaumandreu and Lin (2017) for the case of dynamic pricing, the solution is to augment the production function with a simultaneous estimate of the markups. For example, the system

$$\begin{cases} \ln \frac{R_{jt}}{VC_{jt}} = -\ln(\beta_L + \beta_M) + \ln \mu_{jt} + e_{jt} \\ q_{jt} = \beta_0 + \beta_K k_{jt} + \beta_L l_{jt} + \beta_M m_{jt} + g(h(k_{jt-1}, m_{jt-1}, w_{jt-1} - p_{jt-1}, p_{Mjt-1} - p_{jt-1}, \mu_{jt-1}) + \xi_{jt} + e_{jt}, \end{cases}$$

where the markups μ_{jt} can be specified as flexible as it is convenient, can provide simultaneous consistent estimates of the elasticities of the inputs and markups. Our example features a CD production function, but there is no difficulty in using any other production function (for example with varying elasticities). Estimation can be carried out simultaneously or in two steps. The first equation alone cannot identify separately elasticities and markups, but it can provide a function of both to be imposed as restriction in the second equation.

Appendix A

Let's put an example to show the likely bias. Here we drop firm and time suscripts. Production function is $q = \beta l + \omega + e = q^P + e$ and demand function $q = -\eta p + \delta$. From cost minimization it turns out that $\omega = q^P - \beta l$ so substitution gives $\phi(l, p, \delta) = \beta l + q^P - \beta l$ $= -\eta p + \delta$. Assume that $E(\delta|l, p) = \gamma l$ (we assume no relation between δ and p for simplicity) and write $\delta = \gamma l + r$ in error form. So $\widetilde{\phi}(l, p) = -\eta p + \gamma l$. The true equation underlying second ACF stage is $q = \beta l + g(-\eta p_{-1} + \gamma l_{-1} + r_{-1} - \beta l_{-1}) + \xi + e$. To put a sign to the bias consider that $g(\cdot)$ is autoregresive of known parameter ρ and that variables l_{-1} and p_{-1} are uncorrelated. The equation becomes $q = \beta l - \rho\eta p_{-1} + \rho\gamma l_{-1} + \rho r_{-1} - \rho\beta l_{-1} + \xi + e$. A standard IV estimation for β would be based on the moment $E(l_{-1}q) = \beta E(l_{-1}(l - \rho l_{-1})) -$ $\rho\eta E(l_{-1}p_{-1}) + E(l_{-1}v)$ where $v = \rho\gamma l_{-1} + \rho r_{-1} + \xi + e$. The estimator

$$\widetilde{\beta} = \frac{E(l_{-1}q)}{E(l_{-1}(l - \rho l_{-1}))} = \beta + \frac{E(l_{-1}v)}{E(l_{-1}(l - \rho l_{-1}))}$$

is positively biased because $E(l_{-1}v) = \rho\gamma E(l_{-1}^2) > 0$.

References

Ackerberg, D., L. Benkard, S. Berry and A. Pakes (2007), Econometric Tools for analyzing Market Outcomes, in Heckman, J. and E. Leamer eds., *Handbook of Econometrics,* vol. 6A, 4171-4276.

Ackerberg, D., K. Caves and G. Frazer (2015), "Structural Identification of Production Functions," *Econometrica*, 6, 2411-2451.

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

De Loecker, J. (2013), "Detecting Learning by Exporting," *American Economic Journal: Microeconomics*, 5, 1-21.

De Loecker, J., P. Goldberg, A. Khandelval and N. Pavnik (2016), "Prices, Markups and Trade Reform," *Econometrica,* 84, 445-510.

De Loecker, J. and F. Warzynski (2012), "Markups and Firm-level Export Status," *American Economic Review*, 102, 2437-2471.

Doraszelski, U. and J Jaumandreu (2013), "R&D and Productivity: Estimating Endogenous Productivity," *Review of Economic Studies,* 80, 1338-1383.

Doraszelski, U. and J Jaumandreu (2018), "Measuring the Bias of Technological Change," *Journal of Political Economy*, 126, 3, 1027-1084.

Hall, R.E. (1990), "Invariance Properties of Solow's Productivity Measure." In *Growth/Prod./Unemp.: Essays to Celebrate Bob Solow's Birthday*, Diamond, P. ed., MIT Press.

Jaumandreu, J. and S. Lin (2017), "Pricing Under Innovation: Evidence from Spanish Manufacturing Firms," mimeo, Boston University and Xiamen University.

Jaumandreu, J. and H. Yin (2018), "Comparing Productivity when Products and Inputs Differ in Quality: China Manufacturing Growth 1998-2013," mimeo, Boston University.

Klette, T. and Z. Griliches (1996), " The Inconsistency of the Common Scale Estimators when Output Prices are Unobserved and Endogenous," *Journal of App.Econometrics*, 11, 343-361.

Levinsohn, J. and A. Petrin (2003), "Estimating Production Functions Using Inputs to Control for Unobservables," *Review of Economic Studies*, 70, 317-341.

Olley, S. and A. Pakes (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263-1298.

Peters, B., M. Roberts, V.A. Vuong and H. Fryges (2017), "Estimating Dynamic R&D Choice: An Analysis of Cost and Long-Run Benefits," *Rand journal of Economics,* 48, 409-437.

Solow, R.M. (1957), "Technical Change and the Aggregate Production Function," *The Review of Economics and Statistics*, 312-320.