Endogenous input quality variation as a help to estimate the production function, not a nuisance[*]

Jordi Jaumandreu[†]
Boston University and CEPR

This version: June 2018

First draft: May 2016

Abstract

We discuss how quality varying inputs should be specified in microeconomic production functions and how to use the available information on their prices or treat its absence. OP/LP type of methods to deal with simultaneity allow for a new look at the problem. When output and inputs differ in quality across firms the production function in term of physical quantities is mispecified and quality survives in the apparent unobserved productivity. Consistent coefficient estimates and unbiased estimation of productivity imply to remove unobserved quality from the random disturbances. We find that a general good alternative is specifying the labor input in quantity (number of workers, hours of work) and include in the demand the firm-level average wage as its price. This leaves labor quality in unobserved productivity, something that can be corrected more or less completely according to the available wage information. We also find good support for the solution of deflating the expenditure on materials by an industry index, what is close to measure the input according to its quality, and include the suitable dummies if some exogenous variation is suspected (e.g. regional).

[*] I thank Uli Doraszelski and Heng Yin for useful discussions.

[†] Department of Economics, 270 Bay State Road, Boston, MA02215; Email: jordij@bu.edu.

1. Introduction

Researchers estimating production functions often cannot observe the input prices paid by the firm. However, these prices are likely to vary by exogenous reasons (e.g. geographical area) and/or endogenous choice of the quality of the variety of the inputs to be used. Sometimes what is observed is the physical quantity of the inputs (e.g. number of workers or hours of work), sometimes the expenses on the input or a bundle of inputs (e.g. wage bill or materials bill), and sometimes both. When both physical quantity and expense are observed (e.g. workers and wage bill) a firm-level average price of the input can be computed. But many times only an index of the input price at the industry level is available. And in a few data bases indices at the firm-level. When and how is possible consistent estimation of the parameters of the production function and unbiased estimation of the distribution of unobserved productivity under these different situations of limited information? This note is dedicated to give an answer to this question.

The rest of the note is organized as follows. Section 2 discusses the meaning of estimating production functions under output and input quality variation. Section 3 takes a quick look at the related literature. Sections 4 and 5 establish how to deal with quantities when quality is endogenous. The first sets the framework. The second analyzes estimation. Section 6 explores an alternative route, the estimation of the standarized quantity of the input. Section 7 discuses the special case in which partial information on prices is available. Section 8 contains the conclusions of the note.

2. Output and input quality

The production function describes the frontier of the set of output quantities which can be obtained with all possible combinations of the inputs. Productivity analysis typically considers that this frontier is common to a set of firms except for Hicks neutral differences in its position, i.e. shifts that leave unchanged all relative marginal productivities. These displacements are also considered to be the form of productivity growth of firms over time. If we want to asses without ambiguity the differences in the position of the frontier we must first ensure that outputs and inputs are homogenous and measured in a common way.[1]

If the inputs of a given firm have higher productivity than the inputs of the rest of the firms, without further information we cannot identify which part of the likely relative forward position of the frontier is due to this circumstance. If the output of a firm has higher quality and requires more inputs to produce each unit we cannot identify by how much the frontier lies below other frontiers because this factor. The inherited theory of production functions and measurement of efficiency (across firms, over time) requires homogeneity of output and input.

But production functions are being profusely and increasingly used in the analysis of markets with differentiated products, where outputs differ from firm to firm and inputs are also presumably quite different in their productivity (as they are in the prices that firms pay by them). Hence a tool developed for the analysis with homogeneous outputs and inputs is applied to a quite different situation. This fact, early recognized in the estimation of production functions, has forced researchers to adapt theory and data. It has also generated many discussions on how to deflate expenditures and use the available prices.

In the absence of a theory of production for differentiated products with differentiated inputs there is something that can always be done: to apply the received theory to products measured in units of "equivalent" content, trying to correct systematically for the unwanted effects of heterogeneity. This is a sensible way to retain the properties of substitutability between inputs.[2] The approach has obviously many limitations (a high quality product may be at the end of the day impossible to be produced with a more intensive use of the inputs that produce the low quality version and conversely) but it is likely to give very sensible results if the differences in output and input are non drastic.

[1] An alternative approach allows for differences in the frontier (across firms and time), modeling production functions with random coefficients. In this case the assessment of comparative efficiency must involve comparisons of the shift and the own frontier changes. If one gives economic meaning to the changes identification has to rely on homogeneous input and outputs too.

[2] Some theories of production drop input sustitutability and impose a technology that implies a ladder of ouput and input quality. See, for example, Kremer (1993), Verhoogen (2008) and Kluger and Verhoogen (2011).

A general common production function F, able to compare efficiency across firms with different outputs and inputs needs to have the form (here we drop for simplicity firm and year subindices)

$$h(Q, \alpha) = F(g_i(X_i, \theta_i), ...) \exp(\widetilde{\omega}^* + e),$$

where Q is the quantity of output and α is an index (scalar or vectorial) of product differences and the $g_i(X_i, \theta_i)$ are functions representing amounts of the inputs $i = 1...n$ measured in units of equivalent productive content across firms (by means of combining the indices θ_i of input differences with the physical quantities X_i). When $\alpha = \theta_i = 0$ outputs and inputs are equivalent across firms and time and the relationship applies to physical quantities

$$Q = F(X_i, ...) \exp(\widetilde{\omega}^* + e),$$

but using physical quantities when output and inputs are differentiated amounts to an specification error.

The analysis of both dimensions (input and output quality) can in principle be separated and this is what we will do in this note to focus in input specification. Function $h(\cdot)$ gives the combinations of output quantity and quality attainable with a given amount of the inputs. Notice that the function is separable in quality and the inputs. A simple convenient implementation is for example $h(Q, \alpha) = Q \exp(\alpha)$. In this case the production function can be rewritten as $Q = F(g_i(X_i, \theta_i), ...) \exp(\widetilde{\omega}^* - \alpha + e) = F(g_i(X_i, \theta_i), ...) \exp(\omega^* + e)$, and unobserved productivity turns out to be gross productivity minus output-quality variations (productivity of a given amount of input is smaller the greater output quality). In the rest of this note we assume that the relevant unobserved productivity is net of quality (i.e. $\omega^* = \widetilde{\omega}^* - \alpha$).

3. Literature

The control for quality differences in the inputs at the time of estimating production functions, and the link between these differences and input prices (available and unobserved) has recently been the object of attention in some "structural" estimations. Dorazelski and

Jaumandreu (2013, 2018) experiment with the correction of the labor input for worker skills and De Loecker, Goldberg, Khandelval and Pavcnik (2016) implement a control for unobserved input price variation induced by quality. The need for corrections and the problems involved in the use of input prices are topics that have been around for a while (see below).

What is new is that OP-type procedures of estimation have propiciated another look at the problems. At least in two senses. First, once that the simultaneity problem has been addressed research can get back to focus the effects of old ancillary questions. Second, and more important, the tools to control for simultaneity interact with the old problems offering new solutions.

Input quality is as old as the analysis of productivity. Griliches (1957), one of his first papers, focused on the biases introduced by measuring inputs without taking into account quality differentials. Calling qL to the "true" measurement of labor input, where L is quantity in time units and q a mulitplier that transforms it in "equivalent effective labor units", the analysis goes over the biases when one specifies labor by observed hours, "as if one man's labor was the same as another". Griliches (1964) pushes the topic further by showing how the labor input in farms accompanied by a measure of the level of education improves the fit of the production function and reduces the size of unexplained productivity. Jorgenson and Griliches (1967), uneasy with the size of the Solow residual for the US economy, spent lots of energy in an specification of the inputs that controls, among other things, for changes in quality (see Heckman, 2005, on the contributions of Griliches to quality adjusted input measures).

These efforts receded by two main reasons. First, the reductions in the size of the residual never were drastic. This is also the conclusion of a recent paper by Fox and Smeets (2011) using detailed microdata on the labor input. Second, simultaneity was isolated as the main cause for important potential biases in the estimation of production functions. The availability of panel data revealed both the importance of the simultaneity bias and suggested ways to control for it.

In the new analyses unobserved productivity ω is recognized as highly correlated with the input choices. The exploration of the ways to avoid the simultaneity bias in Griliches and Mairesse (1998) (that includes a summary of OP as an "interesting new approach") contains at least two relevant developments in relation to our topic. First an enlightening and often quoted discussion about the possibility of using input prices as instruments.[3] The paper puts the example that wages are likely to be correlated with cross-section quality differences and time cyclical movements, what makes them unsuitable instruments. But second, much less stressed in the literature, the assessment of the key theoretical role that prices if available play in the identification of the model (page 189 and footnote 40). See also Griliches and Mairesse (1984). So the unsuitability of wage as instrument is seen a consequence of the absence of corrections for labor that make labor quality (and cycle) to be also in unobserved productivity. The correction in itself is no longer too much emphasized because cannot solve the simultaneity bias problem by itself (in unobserved productivity there is more than unobserved quality).

Some recent papers have limited themselves to state that the condition for input prices to be a legitimate instrument is to reflect "exogenous" variation without discussing how this can happen or how the researcher may make this happen.

Ackerberg, Caves and Frazer (2015), henceforth ACF, make clear that their discussion on identification assumes "an “ideal” data scenario where all variables are measured in physical units that are equivalent across firms". Pages 2435 and 2436 offer then a detailed discussion of the implications. If firms are price takers and ouput and input prices are equal for all of them the use of monetary magnitudes is enough. If firms are price makers and it can be assumed that all demand/supply curves and the price choices on them are equivalent for all firms we are in a similar case. But if prices are different, either with price takers or price makers, price variation must be observed an included in the analysis.

[3] "Why do wages differ across firms at a point in time and whitin firms over time? The first is likely to be related to unmeasured differences in the quality of labor, and the second may fluctuate with unexpected shifts in the amount of overtime compensation. Even if there are aggregate time-series movements in real wage and also valid regional wage differences, the use of time and firm dummies (i.e. estimating within or in differences) will eliminate their contribution from the data, leaving us mostly with innapropriate or erroneous variation in these numbers"

The practitioner, who has been attracted in the introduction of the paper by the statement that OP/LP methods "do not require the econometrician to observe exogenous, across-firm variation in input prices" (page 2416), realizes now that this refers exclusively to the case in which it can be assumed that there is no such variation. But the biggest puzzle comes when she re-reads pages 2421-2422, and footnotes 3 and 12, realizing that the article emphasizes that, for being used, the variation in prices should be exogenous and that, in particular, price differences across firms reflecting quality differences are not legitimate.[4] The practitioner, who has just access to some varying prices as firm-level average wages and maybe some materials firm-level prices or price indices, all of them likely affected to some extent by quality differences, is left without alternative.

Confronted with the factual importance of the variation in quality De Loecker, Goldberg, Khandelval and Pavcnik (2016) offer a solution when quality differences are part of the price variation and these price differences are not observed. This alternative doesn't offer again any solution to the problem of how to use the information on prices that is available. The starting point is, like in ACF, a production function in physically equivalent units. After citing OP/LP and the approach in ACF the paper continues: "if we theoretically had data on the physical inputs (...) for all products, these existing approaches to estimating production functions would, in principle, suffice to obtain consistent estimates of the production function coeffients β".

The authors do not find any problem in the output measurement, that is available in physical units, but put forward that the use of "deflated input expenditures" in place of physical quantities introduces an "input bias" (a detailed exposition of the argument of why is also offered in De Loecker and Goldberg, 2014). The paper argues that the bias can be redressed by including in the production function a function of the "output price, market share and product dummies" as control. The idea is that input prices should be present in the specification and that input prices are a function of input quality and exogenous

[4] In fact, the article also explains that with exogenous variation in all prices "estimating the production function using input price based IV methods might be preferred to OP/LP related methodology (due to fewer auxiliary assumptions)."

variation (geography), that input quality can be written as a function of output quality, and that output quality is a function of output price, market shares and product characteristics.

The main problem of this approach is to consider that the right specification of the production function, when there are differences in quality in output and input, is in terms of physical quantities. As explained above, this constitutes mispecification. A physical quantity of output leaves output quality in unobserved productivity. The target for the measures of the inputs is not physical quantities but equivalent quality units, that can be possibly reached by an industry deflator. If there are also exogenous geographical differences they can be controlled by region dummies. According to this the idea of the inclusion of a function to proxy for individual input prices is unnecessary and, in the best case, may be having a completely different effect than interpreted (control of output quality?)[5]

4. Profit maximization and input price variation.

Profit maximization.

Firm j $(j = 1...N)$ has production function

$$Q_{jt} = F(L^*_{jt}) \exp(\omega^*_{jt} + e_{jt}), \tag{1}$$

where $L^*_{jt} = L_j(L_{1jt}, L_{2jt}...L_H)$ is an aggregate variable input, $L_j(\cdot)$ is linearly homogeneous with normalization $\frac{\partial L^*_{jt}}{\partial L_{1_{jt}}} = 1$ (or equivalent if $L_j(\cdot)$ is non-differentiable), ω^*_{jt} represents productivity unobserved by the econometrician and e_{jt} an uncorrelated measurement error. The arguments $L_{1jt}, L_{2jt}...L_H$ of $L_j(\cdot)$ are physical quantities of H varieties of the input.[6] Varieties other than 1 are non-essential, so quantities may be zero. Let's call H_j the set of

[5] The justification of the function is in itself quite problematic. With several inputs, and absent the restrictive Leontieff assumption that characterizes the O'Ring theories of production, the relative cost of each input quality dimension is relevant and should be included in the specification. The exact dimensions of the variable market share and product characteristics and their relation with the output demand assumptions are never clarified. The constant marginal cost used to derive the function contradicts the generality of the specification of the production function.

[6] We do not set any particular functional form restriction on $L_j(\cdot)$. The input varieties may be imperfect substitutes, perfect substitutes or perfect complements. Note that we consider this function firm-specific.

categories employed by the firm. It is straightforward but cumbersome to allow fixed inputs and other variable inputs as arguments of $F(\cdot)$.

Varieties exhibit different quality (marginal productivity) and prices. A natural example is workers (or hours) of different skills and market wages, so we will refer henceforth to this case without loss of generality. We can think of workers of type 1 as some basic category.

If the function is differentiable, by Euler theorem

$$L^*_{jt} = \sum_h \frac{\partial L^*_{jt}}{\partial L_{h_{jt}}} L_{hjt} = L_{jt}\frac{\partial L^*_{jt}}{\partial L_{1_{jt}}}(1+\sum_{h\neq 1}\frac{\frac{\partial L^*_{jt}}{\partial L_{h_{jt}}} - \frac{\partial L^*_{jt}}{\partial L_{1_{jt}}}}{\frac{\partial L^*_{jt}}{\partial L_{1_{jt}}}} s_{hjt}) = L_{jt}(1+\theta_{jt}),$$

where L_{jt} are total workers, $s_{hjt} = \frac{L_{hjt}}{L_{jt}}$ is the share of workers of type h in total employment, and $\theta_{jt} = \sum_{h\neq 1}(\frac{\partial L^*_{jt}}{\partial L_{h_{jt}}} - 1)s_{hjt}$ is an index of quality (in this particular case embedded in skills).[7]

Assume perfect competition for simplicity. If the function is differentiable, firms maximize profits (with expectation $E(\exp(e_{jt})) = 1$)[8] according to the FOCs[9]

$$\begin{aligned} P_t\frac{\partial Q_{jt}}{\partial L^*_{jt}} &= \overline{W}_t \;\text{ for } h=1 \\ P_t\frac{\partial Q_{jt}}{\partial L^*_{jt}}\frac{\partial L^*_{jt}}{\partial L_{hjt}} &= W_{ht} \;\text{ for } h \subset H_j. \end{aligned}$$

Subtracting the two sides of the first condition from the two sides of each one of the others and dividing the result by the two sides of the first, weighting by the shares of each worker type and adding up we get

$$\sum_{\substack{h\neq 1 \\ h\subset H_j}} (\frac{\partial L^*_{jt}}{\partial L_{h_{jt}}} - 1)s_{hjt} = \sum_{\substack{h\neq 1 \\ h\subset H_j}} (\frac{W_{ht} - \overline{W}_t}{\overline{W}_t})s_{hjt}$$

or

$$\theta_{jt} = \sum_{\substack{h\neq 1 \\ h\subset H_j}} (\frac{W_{ht} - \overline{W}_t}{\overline{W}_t})s_{hjt}.$$

[7] If $L_j(\cdot)$ is Leontieff, $L^*_{jt} = \min\{\alpha_h L_{hjt}\}$ with $\alpha_h = 1$, at the minimum we can similarly write $L^*_{jt} = L_{jt}\frac{1}{1+\sum_{h\neq 1}\frac{1}{a_h}}$, and hence $\theta_{jt} = -\frac{\sum_{h\neq 1}\frac{1}{a_h}}{1+\sum_{h\neq 1}\frac{1}{a_h}}$.

[8] More in general assuming that e_{jt} is iid $E(\exp(e_{jt})) = \mu$. We avoid this for the simplicity of notation but recall that this expectation will be, in general, in the constant of the equations.

[9] We denote W_{1t} by $\overline{W}_t$ to emphasize that is the wage for workers of a category that we take as base.

Profit maximization implies that the index of quality is identical to the aggregate of wage premiums.[10] If $L_j(\cdot)$ is Leontieff, we need the market equilibrium to ensure a similar result.

We have started this note saying that available information often consists of the firm-level wage bill WB_{jt}, the number of worker (hours) L_{jt} or both. Noting the above equality, we can write the firm-level average wage (the wage bill WB_{jt} divided by the number of workers or hours) as a function of the salary of the basic category and the index of quality:

$$\frac{WB_{jt}}{L_{jt}} \equiv W_{jt} = \sum_h W_{ht} s_{hjt} = \overline{W}_t + \sum_{h \neq 1} (W_{ht} - \overline{W_t}) s_{hjt} = \overline{W}_t (1 + \theta_{jt}).$$

Notice that the only two key assumptions for this result are profit maximization and that firms face the same set of wages for workers of different qualities in labor markets. The result is robust in particular to different functional forms of the aggregator of varieties of the input and to imperfect competition in the product market.[11] Quality is however going to be mixed with other factors if the firm has monopsony power in the labor market or wages are bargained at the firm level. [12]

By making function $L_j(\cdot)$ firm-specific we allow the greatest flexibility in how firms build their labor input. We are, for example, intentionally vague about why and how the different varieties of workers enter in the set chosen by a particular firm because we do not need this detail. We can think of either production functions with non-essential inputs and input availability which differs from firm to firm (e.g. some worker types are not available in some area markets) or that firms chose among a menu of production functions according to fixed costs of some inputs. The important consequence is that each firm may choose a different optimal worker mix and both the quality of the observed quantity of input L_{jt} and its observed firm-level average price W_{jt} will differ. We summarize this saying that variation in the input and its price is driven by endogenously determined quality differences.

Input price variation: quality-related and other.

Let's address here two important details. First, wages may also differ by exogenous

[10] If $L_j(\cdot)$ is linear (perfect substitition), notice that these FOCs doesn't determine the proportions.

[11] To see this it is enough to replace P_t by the relevant firm-specific marginal revenue.

[12] For example under similar monopsony power in all input varieties characterized by an elasticity of wage with respect to the quantity equal to ε_{jt} the latest formula becomes $W_{jt} = \overline{W}_t(1 - \varepsilon_{jt})(1 + \theta_{jt})$.

reasons. For instance, wages can differ by regions. Second, input quality differences are likely to be related to different output quality. In fact it seems quite natural to think of availability and/or fixed costs of inputs as related to the particular version of the product to be produced.[13]

To formalize the first issue let's assume that the N firms are located in different geographical areas.[14] If the basic category wage $\overline{W}_t$ and the rest of wages were the same everywhere, the different firm-level wages would only express the different quality of workers hired by firms. But assume now that the basic wage is region specific so that we have $\overline{W}_{Rt}$ and, for simplicity, that the rest of regional wages keep the same structure with respect to the specific basic wage everywhere. Now the observed wage for firm j in a given region R is $W_{jt} = \overline{W}_{Rt}(1 + \theta_{jt})$, which continues being an index of the relative quality of the input in the region but it is no longer a pure quality index compared with the wages of firms in other regions. Calling λ_{Rt} the wage premium of a particular region with respect to the average of the regions we can also write $W_{jt} = \overline{W}_t(1 + \lambda_{Rt})(1 + \theta_{jt})$. The differences in wages now include an exogenous variation component by regions.

Our discussion will proceed in the next sections with two cases: when the wage reflects pure endogenous quality differences ($\lambda_{Rt} = 0$) and when there is also an exogenous component ($\lambda_{Rt} \neq 0$).

5. Estimating with input quantities under endogenous quality

Assume now that $F(\cdot)$ is a constant elasticity transformation with elasticity β to facilitate the maths.[15] First notice that when the physical quantity of the input is used in the production function quality becomes a component of unobserved productivity. Using the approximation $(1 + x)^\beta \simeq \exp(\beta x)$ we have

$$\begin{aligned} Q_{jt} &= (L^*_{jt})^\beta \exp(\omega^*_{jt} + e_{jt}) = (L_{jt}(1 + \theta_{jt}))^\beta \exp(\omega^*_{jt} + e_{jt}) \\ &\simeq L^\beta_{jt} \exp(\omega^*_{jt} + \beta\theta_{jt} + e_{jt}) = L^\beta_{jt} \exp(\omega_{jt} + e_{jt}), \end{aligned}$$

[13] Overtime can be seen as a type of labor linked to reach some output quality dimension linked to time.

[14] Other variations of wages can be treated in a similar way. Wages can vary, for example, by subindustry.

[15] The production function is likely to have also a constant that we omit to simplify notation.

where $\omega_{jt} = \omega^*_{jt} + \beta\theta_{jt}$ could be called "apparent" unobserved productivity. In logs we may write

$$q_{jt} = \beta l_{jt} + \omega_{jt} + e_{jt} \tag{2}$$

This simply formalizes an idea that was already advanced in the very first discussions on unobserved productivity (see Griliches, 2000). If one cannot control for quality of the inputs quality becomes part of unobserved productivity.

Second notice that the FOCs corresponding to profit maximization with this production function can be added up and the sum written as

$$P_t\beta(L^*_{jt})^{\beta-1}\exp(\omega^*_{jt})(1+\theta_{jt}) = \overline{W}_t(1+\theta_{jt}). \tag{3}$$

This expression, using the same approximation as before can be transformed into

$$P_t\beta L^{\beta-1}_{jt}\exp(\omega^*_{jt}+\beta\theta_{jt}) = P_t\beta L^{\beta-1}_{jt}\exp(\omega_{jt}) = W_{jt}.$$

The firm demand for (log) workers or hours is hence

$$l_{jt} = \frac{1}{1-\beta}[\ln\beta - (w_{jt}-p_t) + \omega_{jt}]. \tag{4}$$

The number of workers or hours in (2) is “twice” endogenous. Equation (4) shows that the quantity of the input l_{jt} depends on ω_{jt} and hence is correlated with its two components: unobserved productivity ω^*_{jt} and unobserved quality $\beta\theta_{jt}$.

Third, notice that we also have $w_{jt} \simeq \overline{w}_t + \theta_{jt}$. This confirms the conventional wisdom which asserts that the observed average input price is not a legitimate instrument in equation (2): it is of course correlated with the number of workers but it is also correlated with the part of unobserved productivity determined by quality.

An OP/LP method of estimation.

In fact, with persistent unobserved productivity there is a definitive problem to estimate equation (2): w_{jt} is not a valid instrument and the lagged l_{jt-1}, w_{jt-1} and q_{jt-1} variables cannot be used as instruments either (if productivity is persistent they are going to be correlated with ω_{jt} since they are correlated with ω_{jt-1}).[16] There is however an Olley and Pakes (1996) and Levinsohn and Petrin (2003) procedure to estimate the production function. To see this assume that ω_{jt} follows an autoregressive first order exogenous Markov process such that $\omega_{jt} = \rho\omega_{jt-1} + \xi_{jt}$ and set the problem in Doraszelski and Jaumandreu (2013) framework:

$$q_{jt} = \beta l_{jt} + \rho(-\ln\beta + (w_{jt-1} - p_{t-1}) + (1-\beta)l_{jt-1}) + \xi_{jt} + e_{jt}. \tag{5}$$

Setting apart the constant the parameters to estimate in this equation are β and ρ.[17] Under the usual timing assumptions we have at least two available instruments to identify these two parameters: l_{jt-1} and w_{jt-1} (q_{jt-1} could in principle also be used). These are valid instruments because they have been (endogenously) determined before knowing the random shock ξ_{jt}. Hence the specification of the equation using the Markov process assumption makes the parameters identifiable. Since we are applying the markovian assumption to the "apparent" productivity ω_{jt} this deserves a discussion that we continue below.

Notice that if average prices cannot be computed (there is no wage bill to compute average wage) with endogenous quality variation the equation is not identified. Variable l_{jt-1} is then correlated with the unobserved wages and cannot be used as instrument. The extra instrument q_{jt-1} is also of no help for the same reason. This leads to the conclusion in the title: endogenous input quality price variation helps to estimate the production function if we observe this variation and makes it impossible, at least without further assumptions, if we do not observe it.[18]

[16] Here we focus in the correlation of w_{jt} with ω_{jt} because input quality variation but current wage is also likely to be correlated with past values of productivity ω^{*}_{jt} through feedback relationships that the model here doesn't need to specify.

[17] The constant can be set apart for the purposese of identification because is likely to contain other specific parameters in addition of the parameters of main interest. For example, in equation (5) it will include in general the constant of the production function and the expectation of the iid e_{jt} error.

[18] Many papers state routinely nowadays that they follow Levinsohn and Petrin (2003) and rely on an inverted labor (material) demand $l_{jt} = l_t(k_{jt}, \omega_{jt})$ ($m_{jt} = m_t(k_{jt}, \omega_{jt})$) to proxy for productivity (k_{jt} is included as a relevant fixed factor). The above discussion makes clear that the sideline of input prices in these specifications induces inconsistency if endogenous quality choices are present.

Notice that the method of estimation and conclusions do not change if wages are characterized by some exogenous variation too. According to our previous geographical model of wage variation this would simply imply that wage is $w_{jt} \simeq \overline{w}_{Rt} + \theta_{jt}$, but many other forms can also been accommodated. Exogenous wage variation embodied in the variation of firm-level average wages doesn't induce any unobservable and increases the variability of the labor responses. The efficiency of the estimation will be increased.

Caveats.

The previous procedure has two caveats. The first is related to the estimation of productivity ω^*_{jt}, the second to the Markov process assumption. Let's treat them by turn.

It is important to realize that what we recover from the estimation of the parameters is an estimate of the distribution of ω_{jt} (or, more precisely, the distribution of ω_{jt} in differences with respect to its mean). It includes the variation implied by the differences in input quality. Therefore, observing endogenous input quality price variation helps to estimate consistently the parameters of the production function but, as could be expected, cannot provide an estimate of unobserved productivity net of input quality effects.

Getting an estimate of the true productivity ω^*_{jt} is however straightforward if endogenous quality choice may be assumed the *unique* source of variation of wages.[19] Noticing that $\frac{W_{jt}}{\overline{W}_t} = (1 + \theta_{jt})$ and having a β estimate one can subtract ex-post the variation $\beta\theta_{jt}$ from ω_{jt} to get an estimate $\widehat{\omega}^*_{jt} = \widehat{\omega}_{jt} - \widehat{\beta}\theta_{jt}$.

Things are not as easy, however, if the wage variation cannot be excluded to have some exogenous sources. Following up our geographical example it would happen that $\frac{W_{jt}}{\overline{W}_t} = (1 + \lambda_{Rt})(1 + \theta_{jt})$ or, approximating the relationship in logs, $w_{jt} = \overline{w}_{jt} + \lambda_{Rt} + \theta_{jt}$. With some information on skill composition of the labor force and regional (or other exogenous variation sources) dummies one can try to get an estimate of θ_{jt} and perform a correction as before.

We have remarked that the assumption that $\omega_{jt} = \omega^*_{jt} + \beta\theta_{jt}$ follows a first degree Markov Process plays a key role in ensuring consistent estimation. The assumption can be stated

[19] Remember, however, that we may have also an effect of output quality in ω^*_{jt}.

formally as

$$P(\omega^*_{jt} + \beta\theta_{jt} | \omega^*_{jt-1}, \theta_{jt-1}) = P(\omega^*_{jt} + \beta\theta_{jt} | \omega^*_{jt-1} + \beta\theta_{jt-1}),$$

or that productivity including the part determined by quality can be predicted from the previous period values up to an independent disturbance. Since the index of quality is likely to be quite persistent (firms tend to produce with an idiosyncratic quality, that is likely to change slowly), this seems a reasonable assumption.

Suppose, however, that this is not true and it is only true unobserved productivity what follows a Markov process. In this case equation (2) should be rewritten as

$$q_{jt} = \beta l_{jt} + \rho\omega^*_{jt-1} + \xi_{jt} + \beta\theta_{jt} + e_{jt},$$

and since

$$\omega^*_{jt-1} = -\ln\beta + (w_{jt-1} - p_{t-1}) + (1-\beta) l_{jt-1} - \beta\theta_{jt-1},$$

we finally have an expression different from (5)

$$q_{jt} = \beta l_{jt} + \rho(-\ln\beta + (w_{jt-1} - p_{t-1}) + (1-\beta) l_{jt-1}) + \beta(\theta_{jt} - \rho\theta_{jt-1}) + \xi_{jt} + e_{jt}. \quad (6)$$

Now there is an extra term which depends on unobserved quality and is correlated with l_{jt}, l_{jt-1} and w_{jt-1}.

To get an idea of the determinants and magnitude of the bias introduced by this extra term we can do the following simplified analysis. Assume that we know ρ and that quality is a fixed effect ($\theta_{jt} = \theta_{jt-1} = \theta_j$) uncorrelated with the level of standarized employment. The equation becomes

$$q'_{jt} = q_{jt} - \rho(w_{jt-1} - p_{t-1}) - \rho l_{jt-1} = cons + \beta(l_{jt} - \rho l_{jt-1}) + \beta(1-\rho)\theta_j + \xi_{jt} + e_{jt}.$$

Since $l_{jt} - \rho l_{jt-1} = l^*_{jt} - \rho l^*_{jt-1} - (1-\rho)\theta_j$, it is not difficult to show that

$$p\lim\widehat{\beta} = \beta(1 - bias), \text{ with } bias = \frac{V((1-\rho)\theta_j)}{V(l_{jt} - \rho l_{jt-1})}.$$

If growth is independent of the size of the firm we have

$$bias = \frac{V(\theta_j)}{V(\theta_j) + V(l^*_{jt-1}) + V(\Delta l^*_{jt})/(1-\rho)^2}.$$

If quality does not represent a big part of the variations in employment, the variance of employment growth is important enough, and productivity is persistent, the bias is going to be negligible. For example, with all variances equal and $\rho = 0.9$ the bias is less than -1%. Fixed quality could be addressed with fixed effects estimation, and varying quality pseudodifferencing the equation. The likely size of the biases does not seem to justify such approaches.

6. An alternative: estimating the input

Until here we have discussed what happens if we include in the production function the physical quantity of the input L_{jt}. A different available route is to estimate the production function employing an estimate of L^*_{jt}, the quality-corrected amount of the input or amount in standard quality units. Let us first discuss the properties of this kind of estimation assuming that we have a measurement of L^*_{jt} and then examine the possibilities and caveats which surround the estimation of L^*_{jt}.

We can use the aggregate of FOCs (3) to derive the demand for L^*_{jt} :

$$l^*_{jt} = \frac{1}{1-\beta}[\ln\beta - (\overline{w}_t - p_t) + \omega^*_{jt}].$$

Hence, having an estimate of L^*_{jt} allows to set an equivalent to equation (5) as

$$q_{jt} = \beta l^*_{jt} + \rho(-\ln\beta + (\overline{w}_{t-1} - p_{t-1}) + (1-\beta)l^*_{jt-1}) + \xi_{jt} + e_{jt}. \tag{7}$$

Expression (7) shows that the estimation of L^*_{jt} drops from the equation the unobserved endogenous quality variation and its impact on the price. In principle the availability of the basic salary $\overline{w}_t$ and an estimate of l^*_{jt-1} should be then enough to consistently estimate the parameters ρ and β. In addition the approach brings a big advantage: it generates an estimate of productivity ω^*_{jt} isolated from the quality effects.

Identification seems however weaker. The basic salary may not change too much over time and/or the panel may be short and parameter ρ difficult to estimate. This specification also hinders identification when a change over time because other reasons is present in the equation. For example, one could want to specify an in-homogeneous Markov process as $\omega_{jt} = \beta_t + \rho\omega_{jt-1} + \xi_{jt}$.

Olley an Pakes (1996) used this alternative in their labor specification. Despite having both the wage bill WB_{jt} and the number of workers L_{jt} of the companies they did not employ the observed average wage in the estimates. Instead, they divided the wage bill by an industry wage benchmark, taking the result as the estimated amount of labor. In terms of this note, if endogenous quality variation is the *unique* source of wage variation and one divides WB_{jt} by the right reference one gets L^*_{jt} :

$$\frac{WB_{jt}}{\overline{W_t}} = \frac{\sum\limits_h W_h L_{hjt}}{\overline{W_t}} = L_{jt}(1 + \theta_{jt}) = L^*_{jt}.$$

It is important to understand that this solution does not depend on having a measurement of the "level" of the benchmark (as is usually the case for wages). Assume that all what we have is an industry wide index I_t that gives the evolution (but not the level) of $\overline{W_t}$. That is, $\overline{W_t} = I_t\overline{W}_0$ if the index has base year at $t = 0$ ($I_0 = 1$). Deflating the input expenses by this index gives L^*_{jt} up to a constant: $\frac{WB_{jt}}{I_t} = \overline{W}_0 L^*_{jt}$. In the example of production function that we are using this only implies a change of the constant. With more general production function specifications this will amount to have one or more coefficients normalized by the constant $\overline{W}_0$.

This discussion has direct implications on the best form to treat materials. Materials cost is often the only expense recorded of other inputs than labor and represents the value of a broad bundle including expenses in things such as sundry materials, outsourcing of parts and pieces, energy, hired external services and so on. There is no benchmark price available to divide the expenditure. The best approximation possible to quantities is a quantity index, obtainable by a price index with unit value some base period. It turns out that dividing expenditures by the price index is enough to control for differences in endogenous quality under the assumption that the price index evolves as the bundle of benchmark prices would evolve. This is all what one needs to do with materials expenditure for consistency of the production function estimation under the assumption that all price variation is due to endogenous quality choice.

Caveat.

Of course it can be some exogenous input price variation, for example across geographical areas. What happens if this is the case? To explore this first notice that the estimate $\widehat{L}^*_{jt}$ contains a bias,

$$\widehat{L}^*_{jt} = \frac{WB_{jt}}{\overline{W}_t} = \frac{WB_{jt}}{\overline{W}_{Rt}} \frac{\overline{W}_{Rt}}{\overline{W}_t} = L^*_{jt}(1+\lambda_{Rt}),$$

and that this bias is going to be in ω_{jt} :

$$\begin{aligned} Q_{jt} &= (L^*_{jt})^{\beta} \exp(\omega^*_{jt} + e_{jt}) = (\widehat{L}^*_{jt}/(1+\lambda_{Rt}))^{\beta} \exp(\omega^*_{jt} + e_{jt}) \\ &\simeq (\widehat{L}^*_{jt})^{\beta} \exp(\omega^*_{jt} - \beta\lambda_{Rt} + e_{jt}) = (\widehat{L}^*_{jt})^{\beta} \exp(\omega_{jt} + e_{jt}). \end{aligned}$$

We could assume as before a Markov process for the composite ω_{jt} and this will simply imply a bias in the estimation of productivity. But it seems more natural to think of geographical differences as very persistent while ω^*_{jt} follows a Markov process. In this case the demand for the estimated labor input is

$$\widehat{l}^*_{jt} = \frac{1}{1-\beta}[\ln\beta - (\overline{w}_t - p_t) + \omega^*_{jt} - \beta\lambda_{Rt}],$$

and the equivalent to equation (5) looks somewhat to equation (7):

$$q_{jt} = \beta\widehat{l}^*_{jt} + \rho(-\ln\beta + (\overline{w}_{t-1} - p_{t-1}) + (1-\beta)\widehat{l}^*_{jt-1}) - \beta(\lambda_{Rt} - \rho\lambda_{Rt-1}) + \xi_{jt} + e_{jt}. \quad (8)$$

We have some unobserved variation $\lambda_{Rt} - \rho\lambda_{Rt-1}$ and it is correlated with the estimate $\widehat{l}^*_{jt-1}$, like in an errors in variable problem. The origin of the problem is that by using a unique deflator all differences are erroneously assumed due to quality when this is not true because there is also some exogenous variation in wages. On the other hand, the error is not likely to be quantitatively very important as discussed before. And an important property is that these unobserved exogenous differences can in principle be controlled by including this variation in the equation. All we need is measurements or indicators of the exogenous variables. For example, if $\lambda_{Rt} = \lambda_{Rt-1}$ regional fixed effects fully restore consistency. An advantage of this approach is that gets an unbiased estimation of ω^*_{jt} without further corrections.

This discussion enlightens an alternative way of estimating the input L^*_{jt} without introducing any error. Approximating as before the wage relationship in logs, $w_{jt} = \overline{w}_{jt} + \lambda_{Rt} + \theta_{jt}$, one can try to get an estimate of θ_{jt} and use it to construct an unbiased $\widehat{l}^*_{jt}$. With an unbiased estimator the properties underlined in the previous section apply (see Doraszelski and Jaumandreu, 2013 and 2018, for applications).

The discussion of the current and previous sections shows why the alternative of estimating an aggregate input for which we have scarce detail (materials say) by dividing the firm expenses on this input by a reference basic price or an index that is representative of the evolution of the basic price is a good alternative for consistency. If all variation in the input can be assumed to be due to endogenous quality differences estimation using an OP- type method is inefficient but consistent. If prices are suspicious of some exogenous variation this variation should and can be accounted for to reach consistency

7. An special case: firm-level price indices of the cost of materials

A particular situation emerges when there is available a firm-specific price index referred to the cost of materials. Let's assume that this index embodies both exogenous and endogenous variation and that the base year is $t = 0$ without loss of generality. Here we will adapt the notation. Call the cost of materials MB_{jt} and the unobserved price of the basic benchmark bundle $\overline{P}_{Mt}$. The value of the index is

$$I_{jt} = \frac{(1 + \lambda_{Rt})(1 + \theta_{jt})\overline{P}_{Mt}}{(1 + \lambda_{R0})(1 + \theta_{j0})\overline{P}_{M0}},$$

where λ_{Rt} and θ_{jt} represent again regional price differences and the index of quality of the materials employed by the firm. Using this index to deflate the cost of materials we get

$$\widehat{M}_{jt} = \frac{MB_{jt}}{I_{jt}} = (1 + \lambda_{R0})(1 + \theta_{j0})\overline{P}_{M0}\frac{M^*_{jt}}{(1 + \theta_{jt})}.$$

Letting the production function be $Q_{jt} = (M^*_{jt})^{\beta} \exp(\omega^*_{jt} + e_{jt})$, where M^*_{jt} is the aggregate unobservable input, it is easy to see that the replacement of M^*_{jt} gives approximately $Q_{jt} = cons(\widehat{M}_{jt})^{\beta} \exp(\omega^*_{jt} + \beta\theta_{jt} - \beta(\theta_{j0} + \lambda_{R0}) + e_{jt})$.

From aggregate conditions equivalent to (3) and denoting $i_{jt} = \ln I_{jt}$ is easy to derive the demand for (log) $\widehat{M}_{jt}$

$$\widehat{m}_{jt} = \frac{1}{1-\beta}[\ln\beta - \beta\ln\overline{p}_{M0} - (i_{jt} - p_{jt}) + \omega^*_{jt} + \beta\theta_{jt} - \beta(\theta_{j0} + \lambda_{R0})].$$

If we now assume that $\omega_{jt} = \omega^*_{jt} + \beta\theta_{jt}$ follows a Markov process, the equivalent to equations (5),(7) or (8) is

$$q_{jt} = cons' + \beta\widehat{m}_{jt} + \rho(-\ln\beta + \beta\ln\overline{p}_{M0} + (i_{jt-1} - p_{t-1}) + (1-\beta)\widehat{m}_{jt-1}) - \beta(1-\rho)(\theta_{j0} + \lambda_{R0}) + \xi_{jt} + e_{jt}. \tag{9}$$

Expression (9) makes clear that deflation by the firm-level specific indices induces some fixed effects linked to the unobserved starting level differences in prices (both exogenous and endogenous) correlated with $\widehat{m}_{jt}$, although it also underlines that the impact of these fixed effects is likely to be quantitatively small. The estimated ω_{jt} includes the effect of quality, as when we use the quantity of the input. Possibilities of correction for this fact are discussed above.

8. Concluding remarks

This note shows:

When the unique reason for variation of the input firm-level price is endogenous choice of quality, consistent estimates of the parameters and unbiased estimation of productivity can be obtained by including in the production function the firm-level expenses deflated by an industry index and using an OP/LP method of estimation. The reason is that deflation in this way measures the input in units of standard quality.

When there is also variation of the firm-level price by exogenous reasons the application of the previous method produces inconsistent estimates and a biased estimation of productivity. The reason can be thought of as a problem of error in variables. The most straightforward form to avoid it is including as controls the exogenous variation (e.g. regional input prices). Another is to estimate properly the amount of the input in units of standard quality.

On the other hand, when there is variation in quality the inclusion of the physical quantity of the input in the production function (e.g. number of workers or hours of work) leaves the quality effect in unobserved productivity creating "double" endogeneity of the input (its quantity is both correlated with true productivity and with quality). However, an OP/LP method that uses the computed firm-level average price as explanatory variable in the demand for the input estimates consistently the parameters of the production function and recovers an estimate of productivity that includes the quality effect.

This method is robust to the presence of exogenous variation in prices and more efficient than the previous one, but also more dependent on the markovian assumption. The productivity estimate can be netted-out ex-post of quality if needed. Correction is straightforward if quality is the unique source of wage variation and less direct otherwise.

These results suggest a different treatment of the input labor and of the expenditure on a bundle of materials that is available in most data bases. Recognizing that wages are likely to show both exogenous and endogenous quality variation, the best alternative seems the use of the quantity of work (workers or hours) and the computed average wage to specify the input demand to apply an OP/LP procedure. Under the assumption that material price differences are only endogenous, consistent estimation can be achieved by including the expenditures in materials deflated by an industry index. If there are reasons to think that there is also exogenous price variation the simplest alternative for consistency is to include it.

References

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

De Loecker, J., and P. Goldberg (2014), "Firm Performance in a Global Economy," *Annual review of Economics,* 6, 201-227.

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

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.

Fox, J. and V. Smeets (2011), "Does input Quality Drive Measured Differences in Firm Productivity?," *International Economic Review,* 52, 961-989.

Griliches, Z. (1957), "Specification Bias in Estimates of the Production Functions," *Journal of Farm Economics,* 39, 8-20.

Griliches, Z. (1964), "Research Expenditures, Education, and the Aggregate Agricultural Production Function," *American Economic Review*, 54, 961-974.

Griliches, Z. (2000), *R&D, Education, and Productivity: A Retrospective,* Harvard University Press.

Griliches, Z. and J. Mairesse (1984), "Productivity and R&D at the Firm Level," in Griliches, Z., ed, *R&D, Patents and Productivity,* University of Chicago Press, 339-374.

Griliches, Z. and J. Mairesse (1998), "Production Functions: The Search for Identification," in Griliches, Z., *Practicing Econometrics: Essays in Method and Applications,* Cheltenham, 383-411.

Heckman, J. (2005), "Contributions of Zvi Griliches," *Annales d'Economie et de Statistique,* 79-80, 5-22.

Jorgenson, D. and Z. Griliches (1967), "The Explanation of Productivity Change," *Review of Economic Studies*, 34, 249-283.

Kluger, M. and E. Verhoogen (2011), "Prices, Plant Size and Product Quality," *Review of Economic Studies*, 79, 307-339.

Kremer, M. (1993), "The O-Ring Theory of Economic Development," *Quarterly journal of Economics,* 551-575.

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.

Verhoogen, E. (2008), "Trade, Quality Upgrading and Wage Inequality in the Mexican Manufacturing Sector," *Quarterly Journal of Economics,* 123, 489-530.