Chapter 1: Production function estimation

Joan Llull

Structural Econometrics for Labor Economics
and Industrial Organization

IDEA PhD Program

I. Introduction

Production functions are important elements in Economics. Specifically, they are fundamental in Industrial Organization, and also have many application in Labor Economics. They play a key role in: determining aggregate productivity and its dispersion (firm heterogeneity, wage inequality,...); miss-allocation of resources; estimation of marginal costs, marginal productivity, and input prices; skill-labor intensity and skill-biased technical change; learning-by-doing; technology adoption and endogenous innovation;...

II. Firm-level estimation

The estimation of firm-specific production function has important applications in Industrial Organization and in micro-founded Macroeconomics among other fields. Applications include: the estimation of marginal costs, productivity, and input prices; innovation and technology adoption decisions; firm dynamics; mergers and acquisitions; the analysis of miss-allocation of resources; and many others.

A. Model setup: a simple Cobb-Douglas production framework

Let y_{it} denote output of firm i at time t, and k_{it} and l_{it} the two inputs used in production, namely capital and labor. Consider a simple Cobb-Douglas production technology:

$$y_{it} = \zeta_{it} k_{it}^{\alpha} l_{it}^{\beta}, \tag{1}$$

where ζ_{it} is firm's total factor productivity. Taking logs to this expression yields:

$$\ln y_{it} = \alpha \ln k_{it} + \beta \ln l_{it} + \nu_{it} + \varepsilon_{it}, \tag{2}$$

where $\nu_{it} \equiv \ln \zeta_{it}$ is the productivity term, unobserved by the econometrician, and ε_{it} is measurement error. This equation provides an example that even a linear regression can be a structural model, if its parameters have an interpretation in terms of the parameters of an economic model.

The linear regression estimation of (2) entails one of the most classic examples of endogeneity discussed in econometrics: the simultaneity bias. This bias arises because the firm knows ν_{it} when deciding the quantities of inputs k_{it} and l_{it} to be used in production. To illustrate it, consider the firm's maximization problem:

$$\max_{k_{it},l_{it}} \zeta_{it} k_{it}^{\alpha} l_{it}^{\beta} - r_{it} k_{it} - w_{it} l_{it}, \tag{3}$$

where r_{it} and w_{it} are input prices, which leads to the demands:

$$\begin{cases} w_{it} = \beta \zeta_{it} k_{it}^{\alpha} l_{it}^{\beta-1} \\ r_{it} = \alpha \zeta_{it} k_{it}^{\alpha-1} l_{it}^{\beta} \end{cases} \quad \Rightarrow \quad \begin{cases} l_{it}^{1-\alpha-\beta} = \zeta_{it} \frac{\beta}{w_{it}} \left(\frac{w_{it}\alpha}{r_{it}\beta} \right)^{\alpha} \\ k_{it}^{1-\alpha-\beta} = \zeta_{it} \frac{\alpha}{r_{it}} \left(\frac{r_{it}\beta}{w_{it}\alpha} \right)^{\beta} \end{cases}. \tag{4}$$

Another source of bias is given by measurement error in the inputs (especially capital), which leads to attenuation bias if the measurement error is classical. An additional type of bias is the selection bias, because there are endogenous exits of firms: survivor firms are not randomly chosen. For example, survivor firms may have higher productivity and use larger amounts of labor and capital than exiting firms. That is, if d_{it} denotes an indicator variable that equals one if the firm is operating, then $\mathbb{E}[\nu_{it} | k_{it}, l_{it}, d_{it} = 1] \neq 0$.

B. Instrumental variables estimation

To deal with the simultaneity and measurement error biases, one of the classic solutions is to use input prices as instruments. In particular, the demands in (4) are functions of prices and productivity. In a competitive setting, in which firms operate in different markets in a competitive way, there is variation in the input prices they face, but this variation is exogenous to firm's productivity. However, this approach generates some tensions and difficulties. First, input prices may not be observable. Second, if firms operate in non-competitive settings, input prices may be affected by firm's productivity (for example, more productive firms being larger and being able to buy at cheaper prices). Third, in a competitive setting, the presence of variation in prices rejects itself the constant parameter model since$\beta = \frac{w_{it} l_{it}}{y_{it}}$ is not constant in the data.

C. Dynamic panel data

An alternative approach is to consider $\nu_{it} = \eta_i + \delta_t + \upsilon_{it}$, where υ_{it} is unknown by the firm at the time of setting up demands. In this context, if 1) it is plausible that υ_{it} is uncorrelated with inputs demands, 2) υ_{it} is i.i.d. over time, and 3) there is within-firm over time variation in input demands, then a panel data fixed effects (within groups) estimation would lead to consistent estimates of α and β.

However, in most applications, these are restrictive assumptions. In many applications in practice, this estimator provides very small estimates of α and β. There are at least two main reasons for that. First, while the presence of variation in inputs within firms over time and the absence of auto-correlation in the productivity term may be plausible in agricultural firms in developing countries, it is rather unlikely to hold in the manufacturing sector in developed countries. And second, the attenuation bias induced by measurement error is exacerbated by the within groups transformation, especially when there is little variation in inputs.

Many papers in the literature relax these assumptions using dynamic panel data models. Strict exogeneity can be relaxed assuming dynamic demands:

$$k_{it} = F_K(k_{it-1}, l_{it-1}, \nu_{it}) \quad \text{and} \quad l_{it} = F_L(k_{it-1}, l_{it-1}, \nu_{it}). \tag{5}$$

There are multiple reasons why the demand for capital or and labor are dynamic: hiring and firing costs for labor, irreversibility of some capital investments, installation costs, time-to-build,.... If demands are dynamic, k_{it-j}, l_{it-j}, and Y_{it-j} for $j \geq 2$ are valid instruments for the regression:

$$\Delta \ln y_{it} = \alpha \Delta \ln k_{it} + \beta \Delta \ln l_{it} + \Delta \delta_t + \Delta \upsilon_{it}, \tag{6}$$

as long as υ_{it} is i.i.d over time. This assumption is again restrictive, but can be tested (as in Arellano and Bond, 1991). In practice, this assumption is very often rejected. Furthermore, the instruments are often weak (because of the strong persistence in the demands), the estimation of the equation in first differences eliminates the cross-sectional variation and exacerbates the measurement error problem, and it often provides downward biased and imprecise estimates (Blundell and Bond, 1998, 2000).

Blundell and Bond (2000) suggest to modify the model to include persistent errors in the following way: $\nu_{it} = \rho \nu_{it-1} + \eta_i + \delta_t + \upsilon_{it}$. Under this assumption, (ignoring the measurement error term) Equation (2) can be rewritten as:

$$\ln y_{it} = \rho \ln y_{it-1} + \alpha(\ln k_{it} - \rho \ln k_{it-1}) + \beta(\ln l_{it} - \rho \ln l_{it-1}) + \eta_i + \delta_t + \upsilon_{it}. \tag{7}$$

The authors then suggest to use the estimation methods discussed in Blundell and Bond (1998), which is based on the methods proposed by Arellano and Bond (1991) and Arellano and Bover (1995) reviewed in the Microeconometrics course.

D. Control function approaches

Olley and Pakes (1996) and Levinshon and Petrin (2003) propose "control function approaches" as an alternative to finding instruments for l_{it} and k_{it}. Intuitively,

they look for observable variables that can "control" for unobserved total factor productivity. These control variables come from a model of firm behavior.

The Olley and Pakes (1996) considers the following modification of the model presented above. The production structure is given by (2). The dynamic demands are given by the following modification of (5):

$$i_{it} = F_K(k_{it}, l_{it-1}, \nu_{it}, \boldsymbol{r}_{it}) \quad \text{and} \quad l_{it} = F_L(k_{it}, l_{it-1}, \nu_{it}, \boldsymbol{r}_{it}), \tag{8}$$

where i_{it} denotes investment at time t, and $\boldsymbol{r}_{it}$ is the vector of factor prices, in this case, $\boldsymbol{r}_{it} = (r_{it}, w_{it})'$. The authors assume: i) $F_K(\cdot)$ is invertible in ν_{it}; ii) no cross-sectional variation in prices, $\boldsymbol{r}_{it} = \boldsymbol{r}_t$ for all i; iii) ν_{it} follows a first order Markov process; and iv) investment i_{it} is chosen in period t, but it is not productive until $t+1$, when $k_{it+1} = (1-\delta)k_{it} + i_{it}$ (where δ is the depreciation rate). They also assume that labor is a perfectly flexible input, that is, l_{it-1} is not a state variable. Following the discussion in Aguirregabiria (2019), we discuss the case in which the latter assumption is relaxed, allowing for labor adjustment costs (this assumption is innocuous in Olley and Pakes, 1996). This approach deals with the simultaneity problem, and can be adjusted to also deal with the endogenous exit bias, which happens to be important in practice (see Aguirregabiria, 2019).

Olley and Pakes proceed with two steps. In the first step, they estimate β using a control function approach based on the first two assumptions above (invertibility and no cross-sectional variation in prices). In particular, this step estimates:

$$\ln y_{it} = \beta \ln l_{it} + \phi_t(l_{it-1}, k_{it}, i_{it}) + \varepsilon_{it}, \tag{9}$$

where $\phi_t(l_{it-1}, k_{it}, i_{it}) \equiv \alpha \ln k_{it} + F_K^{-1}(l_{it-1}, k_{it}, i_{it}, \boldsymbol{r}_t)$, and where $F_K^{-1}(l_{it-1}, k_{it}, i_{it}, \boldsymbol{r}_{it})$ is the inverse of $F_K(\cdot)$ with respect to ν_{it}. This equation can be estimated without imposing any parametric assumption on $F_K(\cdot)$, which yields to a semi-parametric partially linear model. This model can be estimated using semi-parametric methods like kernel regressions, or approximating $\phi_t(\cdot)$ by means of polynomial series approximations, as in Olley and Pakes (1996). As we discuss below, Ackerberg, Caves, and Fazer (2006) noted that, on top of invertibility and no cross-sectional variation in prices, this step requires that there is enough variation in l_{it} to identify β after controlling for l_{it-1}, k_{it}, and i_{it}.

The second step entails the estimation of α given the estimate $\hat{\beta}$ obtained in the first stage. To this end, we need the additional two assumptions (Markovian nature of ν_{it} and the time-to-build assumption for capital). Since ν_{it} is Markovian:

$$\nu_{it} = \mathbb{E}[\nu_{it}|\nu_{it-1}] + \xi_{it} \equiv h(\nu_{it-1}) + \xi_{it}, \tag{10}$$

where ξ_{it} is a mean-independent innovation and $h(\cdot)$ is some unknown function. Given the definition of $\phi_t(l_{it-1}, k_{it}, i_{it})$ and the identity $\nu_{it} = F_K^{-1}(l_{it-1}, k_{it}, i_{it}, \boldsymbol{r}_t)$:

$$\begin{aligned}\hat{\phi}_{it} &= \alpha \ln k_{it} + h(\nu_{it-1}) + \xi_{it} \\ &= \alpha \ln k_{it} + h(\hat{\phi}_{it-1} - \alpha \ln k_{it-1}) + \xi_{it},\end{aligned} \tag{11}$$

where $\hat{\phi}_{it} \equiv \ln y_{it} - \hat{\beta} \ln l_{it}$. This model is again a partially linear model, as (9). However, unlike in the case of (9), the argument of the $h(\cdot)$ function is not observable, because it depends on the unknown parameter α. Olley and Pakes propose a recursive version of the semiparametric method in the first step. In particular, they start from a guess of α, namely $\alpha^{(0)}$, and compute $\hat{\phi}_{it-1} - \alpha^{(0)} \ln k_{it-1}$. Given this, they obtain a next guess $\alpha^{(1)}$ as the coefficient of $\ln k_{it}$ in the semi-parametric regression (11). If the new guess equals the preceding one, the algorithm stops; otherwise, it proceeds with a new iteration replacing $\alpha^{(0)}$ by $\alpha^{(1)}$, until reaching convergence. Alternatively, one can similarly use a minimum distance estimator.

Levinshon and Petrin (2003) adjust the Olley-Pakes algorithm to account for two important issues: i) investment can be responsive to more persistent shocks in TFP; and ii) zero investment accounts are very present in many data-sets (at $i_{it} = 0$, corner solution, there is no invertibility between i_{it} and ν_{it}). Instead of the primary investment function to generate the control function, they use the demand function for intermediate inputs. Consider the following version of the (linearized) production function:

$$\ln y_{it} = \alpha \ln k_{it} + \beta \ln l_{it} + \gamma \ln m_{it} + \nu_{it} + \varepsilon_{it}, \tag{12}$$

where m_{it} denotes intermediate inputs (materials). The investment equation is now replaced with the demand for materials:

$$m_{it} = F_M(l_{it-1}, k_{it}, \nu_{it}, \boldsymbol{r}_{it}), \tag{13}$$

assumed, again, to be invertible in ν_{it} (in this case, $\boldsymbol{r}_{it} = (r_{it}, w_{it}, p_{it})'$, where p_{it} denotes the price/cost of materials). The assumptions of no cross-sectional variation in prices, first order Markovian TFP, and time-to-build for capital are still assumed to hold. In this case, the first step is analogous to that in Olley and Pakes, except that investment is replaced by the demand for materials:

$$\ln y_{it} = \beta \ln l_{it} + \varphi_t(l_{it-1}, k_{it}, m_{it}) + \varepsilon_{it}, \tag{14}$$

where $\varphi_t(l_{it}, k_{it}, m_{it}) \equiv \alpha \ln k_{it} + \gamma \ln m_{it} + F_M^{-1}(l_{it-1}, k_{it}, m_{it}, \boldsymbol{r}_{it})$, and where the function $F_M^{-1}(l_{it-1}, k_{it}, m_{it}, \boldsymbol{r}_{it})$ is the inverse of $F_M(\cdot)$ with respect to ν_{it}.

The second step is also analogous to Olley and Pakes:

$$\begin{aligned}\hat{\phi}_{it} &= \alpha \ln k_{it} + \gamma \ln m_{it} + h(\nu_{it-1}) + \xi_{it} \\ &= \alpha \ln k_{it} + \gamma \ln m_{it} + h(\hat{\phi}_{it-1} - \alpha \ln k_{it-1} - \gamma \ln m_{it-1}) + \xi_{it}. \end{aligned} \tag{15}$$

This second step, however, includes an important difference with respect to Olley and Pakes: $\mathbb{E}[\xi_{it} \ln m_{it}] \neq 0$ (it is zero for capital because of the time-to-build assumption). In order to account for the subsequent endogeneity of m_{it} they propose to instrument by lagged values of m_{it} in the spirit of Blundell and Bond (see an important critique in Gandhi, Navarro, and Rivers, 2020).

Ackerberg, Caves, and Fazer (2015) provide an important critique (and a solution) to the methods proposed by Olley and Pakes (1996) and Levinshon and Petrin (2003). In the following lines, we discuss their critique focusing on the Olley and Pakes case, but the extension to the Levinshon and Petrin framework is trivial. Given the assumptions of invertibility and no cross-sectional variation in prices, we can rewrite the labor demand as:

$$l_{it} = F_L(l_{it-1}, k_{it}, F_K^{-1}(l_{it-1}, k_{it}, i_{it}, \boldsymbol{r}_t), \boldsymbol{r}_t) \equiv G_t(l_{it-1}, k_{it}, i_{it}). \tag{16}$$

Therefore, once l_{it-1}, k_{it}, and i_{it} are non-parametrically controlled for in (9), there should be no variation left to identify β. Therefore, in practice, either the model is incorrectly specified, or β is identified spuriously.

These authors discuss alternative specifications in which the Olley and Pakes (or Levinshon and Petrin) approach identifies the parameters of interest and provides consistent estimates. The main requirement is some "exclusion restriction" for the labor demand, that is, some variable that generates variation in l_{it} when we hold l_{it-1}, k_{it}, and i_{it} fixed. In particular, they assume that:

$$i_{it} = F_K(k_{it}, l_{it-1}, \nu_{it}, r_{it}) \quad \text{and} \quad l_{it} = F_L(k_{it}, l_{it-1}, \nu_{it}, w_{it}), \tag{17}$$

where the input prices w_{it} and r_{it} satisfy that, conditional on t, i_{it}, k_{it}, and l_{it-1}, i) w_{it} has cross-sectional variation, that is, $\text{Var}(w_{it}|t, i_{it}, k_{it}, l_{it-1}) > 0$, and ii) w_{it} and r_{it} are independently distributed. These assumptions are consistent with the following economic assumptions: a) capital markets are perfectly competitive and the price of capital is the same for every firm, i.e. $r_{it} = r_t$ for all i (this assumption is not strictly necessary, because r_{it} does not enter the labor demand, but the independence of w_{it} and r_{it} is harder to sustain otherwise); b) there are internal labor markets such that the price of labor has cross-sectional variability; c) the realization of the cost of labor occurs after the investment decisions take place,

and, therefore, w_{it} does not enter the investment function; and d) the idiosyncratic labor costs are not serially correlated (so that lagged labor cost shocks are not state variables in the optimal investment decision).

III. Aggregate production functions

Aggregate production functions are a central element for Macroeconomics and for Labor Economics. They are fundamental to describe input demands in general and partial equilibrium models, to identify technological shocks and total factor productivity, and to describe the evolution of input prices (e.g. wages). In this context, the focus is more often placed on the elasticities of substitution across inputs and in the residuals, and the evolution of the technology parameters. The construction/estimation of factor shares and depreciation rates is often approached from a perspective of measurement/accounting.

Within the bulk of research that investigates (or implements) the estimation of aggregate production functions, we can distinguish between those that exploit general equilibrium conditions and those based on partial equilibrium. Estimation within equilibrium frameworks is not particularly convoluted because the explicit modeling of supply and demand explicitly deals with most of the potential endogeneity concerns. In this section, instead, we focus on the estimation of production functions in partial equilibrium frameworks.

In the context of partial equilibrium, we can distinguish between approaches that exploit cross-sectional (typically spatial) variation and those that are based on time series (and, potentially, cross-input) variation. In the latter case, technological/TFP shocks are often specified in a Hicks-neutral way, which often facilitates the estimation of some of the parameters by least squares methods, as we discuss below for the constant elasticity of substitution case.

There is a very extensive literature directly or indirectly focused on the estimation of aggregate production functions that will not be reviewed here. Instead, what we do is to use a few examples to illustrate some of the most common issues faced by researchers estimating aggregate production functions. The application in Section IV below provides an additional example.

A. Elasticities of substitution: nested constant elasticity of substitution

One of the most convenient ways to estimate elasticities of substitution across inputs (or allow for imperfect substitutability across them) is by means of nested constant elasticity of substitution (CES) production functions. The two main advantages of these production functions are: i) they exhibit a log-linear relation

between relative prices and relative inputs, and ii) the elasticity of substitution between two inputs inside one nest can be estimated without information on the inputs or parameters in the nests that lie above the nest of interest.

Consider the following production function, based on the papers by Card and Lemieux (2001), Borjas (2003), and Ottaviano and Peri (2012):

$$Y_t = A_t K_t^{\alpha} L_t^{1-\alpha}, \tag{18}$$

with:

$$L_t \equiv \left[\sum_i \theta_{it} L_{it}^{\rho}\right]^{\frac{1}{\rho}}, \; L_{it} \equiv \left[\sum_j \gamma_{ij} L_{ijt}^{\eta}\right]^{\frac{1}{\eta}}, \text{ and } L_{ijt} \equiv [\lambda L_{ijNt}^{\phi} + (1-\lambda) L_{ijMt}^{\phi}]^{\frac{1}{\phi}}, \tag{19}$$

where i indexes education groups, j indexes experience (age) groups, and N and M denote, respectively, natives and immigrants. Card and Lemieux (2001) estimates this production function (without the lowest level) to account for the importance of imperfect substitutability across cohorts in explaining the increasing wage inequality. Borjas (2003) and Ottaviano and Peri (2012) estimate versions of this production function using data from the U.S. Census in order to simulate the effect of immigration on wages in each education-experience cell. Borjas (2003) ignores the lowest layer, and Ottaviano and Peri (2012) estimate different versions with different nesting orders. With so few periods, the value of α is often assumed out (e.g., to be 0.3). Therefore, the parameters left to be estimated are those associated with the elasticities of substitution across inputs, ϕ, η, and ρ.

The estimation of nested CES production functions is often based on the first order conditions of the (aggregate competitive) firm's optimization problem (that is, wages equal marginal product), and it is implemented sequentially. As noted above, the relative prices (wages) of a pair of inputs included within the same nest does not depend on the demands of inputs in upper levels. In this production function, the relative wages of natives and immigrants with education i and experience j is:

$$\begin{aligned} \ln \frac{w_{ijMt}}{w_{ijNt}} &= \ln \frac{\partial Y_t / \partial L_{ijMt}}{\partial Y_t / \partial L_{ijNt}} \\ &= \ln \left(\frac{\partial Y_t / \partial L_{ijt}}{\partial Y_t / \partial L_{ijt}} \times \frac{\partial L_{ijt} / \partial L_{ijMt}}{\partial L_{ijt} / \partial L_{ijNt}} \right) \\ &= \ln \frac{1-\lambda}{\lambda} + (\phi - 1) \ln \frac{L_{ijMt}}{L_{ijNt}}. \end{aligned} \tag{20}$$

Note that this expression is exact, given that wages and labor inputs are observed

(and the only unobservable of the production function, A_t, cancels out in the first term of the second line of the expression). Adding a measurement error in relative wages to (20), λ and ϕ is obtained from a least squares estimation.

Having identified λ and ϕ in the first stage, the labor inputs L_{ijt} can be constructed. In order to identify the parameters in the following nesting level, η and γ_{ij}, one could proceed analogously, deriving the equivalent expression for every pair of experience groups. Equivalently, these parameters can be estimated in levels, using fixed effects to capture the common terms that would cancel in the relative wage expressions.[1] The wage in cell ijt is given by:

$$\ln w_{ijt} = \ln\left(\frac{\partial Y_t}{\partial L_t} \times \frac{\partial L_t}{\partial L_{it}} \times \frac{\partial L_{it}}{\partial L_{ijt}}\right) = \kappa_t + \pi_{it} + \ln \gamma_{ij} + (\eta - 1) \ln L_{ijt}. \quad (21)$$

Taking into account the potential measurement error in the estimation of wages in cell ijt, the above expression can be trivially estimated by least squares including education-time and education-experience fixed effects (dummies). Given these, η is identified from the coefficient on the labor input variable, and γ_{ij} is identified from the coefficients of the education-experience dummies (up to some normalizations, which usually set $\sum_j \gamma_{ij} = 1$ for every education group i).

Once the second step is completed with estimates of γ_{ij} and η, the labor input in the next nesting level, L_{it}, can be computed. Then, the third step proceeds analogously to estimate ρ and θ_{it}, subject to some normalization on θ_{it} (given that fixed effects would absorb all the available degrees of freedom):

$$\ln w_{it} = \kappa_t + \ln \theta_{it} + (\rho - 1) \ln L_{it}. \quad (22)$$

Borjas (2003) specifies θ_{it} as education group-specific time trends.

In order to account for potential model specification error, potentially correlated with the labor inputs, Borjas (2003) and Ottaviano and Peri (2012) also estimate (21) and (22) using the stock of immigrants in each cell as an instrument for the labor inputs. In this case, the results with and without instrumentation are very similar, which could be interpreted as evidence that the need for instrumentation in this context is limited.

B. The race between technology and skills

Another common application of aggregate production functions in Labor Economics and Macroeconomics is the use of the estimated parameters to account for

[1] The comparison between the estimation in levels or in terms of relative wages is analogous to the estimation of fixed effects models in first differences versus with the standard within-groups estimation.

the determinants of the increasing wage inequality. Acemoglu and Autor (2011) describe the "canonical model" of wage inequality as follows. Consider two skills, H_t (for high) and L_t (for low). Output is produced with the following technology:

$$Y_t = (\alpha_{Lt} L^{\rho} + \alpha_{Ht} H^{\rho})^{\frac{1}{\rho}}. \tag{23}$$

As noted above, relative wages are given by:

$$\ln \frac{w_H}{w_L} = \rho \ln \left(\frac{\alpha_{Ht}}{\alpha_{Lt}} \right) + (\rho - 1) \ln \frac{H_t}{L_t}. \tag{24}$$

Note that we indexed α_{Lt} and α_{Ht} by t to capture technological progress. This expression describes what Tinbergen described as the "race between technology and skills". Intuitively, the relative wages of high and low skilled workers decrease with the increase in the relative supply of high skilled workers (ρ is typically a number between zero and one provided that L_t and H_t are, potentially imperfect, substitutes) and increase if there is skill-biased technical change (α_{Ht}/α_{Lt} grows).

The seminal work by Katz and Murphy (1992) estimate this regression by OLS, assuming that the first term is well captured by a time trend. Note that this assumption is crucial: if $\frac{\alpha_{Ht}}{\alpha_{Lt}}$ is, instead, assumed to be a random variable (included in the error term), it would be correlated with $\frac{H_t}{L_t}$, which would lead to biased estimates of ρ. In this case, the nature of the biases would be similar to those described in Section II. However, the time series nature of the data limits the use of some of the approaches described there. In practice, most papers exploiting national-level time series variation tend to ignore these concerns. Alternatively, approaches that exploit spatial variation, described in Section III.E below, rely on instrumental variable approaches.

This simple model fits the data quite well, at least within sample, until mid-1990s, after which over-predicts the increase in the college-high school wage gap. If the model is enriched to account for non-linear trends (linear spline, quadratic trend, cubic trend, etc.), the model fits the data well for the entire period, but all trends suggest that the relative demand for college workers *decelerated* in the 1990s, which seems counter to the common perception of how the technological progress occurred in this period. Other papers like Card and Lemieux (2001) or Jeong, Kim, and Manovskii (2015) expanded this model to account for different groups of workers. Card and Lemieux (2001) estimate a nested CES production function like the one discussed below showing that the imperfect substitutability between young and old workers within an education group and the changes in the cohort sizes can explain a part of the discrepancy. Jeong, Kim, and Manovskii

(2015) depart from the college-high school classification of skills to distinguish between labor and experience as separate skill inputs. Their framework is useful to illustrate the use of alternative approaches to the linear regression in (24). We review these methods in Section IV, where we review Albert, Glitz, and Llull (2020), who estimate a different production function with similar non-linear methods.

C. Capital-labor substitution and biased technical change

The capital-labor elasticity of substitution is a very important parameter in economics. Many papers assume that it is one (Cobb-Douglas), which implies that the technological progress is Hicks-neutral. Other papers try to test whether it is one or not, making assumptions about the particular structure of the technological progress. Antràs (2004) show that, indeed, in a context of relatively stable factor shares, the assumption of factor neutral technical progress biases the results towards Cobb-Douglas. Diamond, McFadden, and Rodríguez (1978) argue that this elasticity and biased technical change cannot be simultaneously identified.

León-Ledesma, McAdam, and Willman (2010) provide a Monte-Carlo analysis to asses under which conditions these two are well identified and robust. Consider the following CES framework:

$$Y_t = \zeta \left(\pi (\Gamma_{Kt} K_t)^{\rho} + (1-\pi)(\Gamma_{Lt} L_t)^{\rho} \right)^{\frac{1}{\rho}}. \tag{25}$$

The elasticity of substitution between capital and labor is given by $1/(1-\rho)$, the time-varying parameters Γ_{Kt} and Γ_{Lt} denote efficiency (and, its evolution, technical progress), the parameter ζ is an efficiency parameter, and $\pi \in [0,1]$ is the capital intensity. This production function embeds the Cobb-Douglas case when $\rho = 0$, the Lenotieff case when $\rho \to \infty$, and the linear case when $\rho \to 1$.

Functional forms are often imposted to Γ_{Kt} and Γ_{Lt} in order to fix the non-identification problem of Diamond, McFadden, and Rodríguez (1978). One of the common cases is to impose (log-) linear trends $\Gamma_{Kt} = e^{-\gamma_K t}$ and $\Gamma_{Lt} = e^{-\gamma_L t}$. Furthermore, depending on the assumed relation between the two, one can be imposing Hicks-neutral technical change ($\gamma_K = \gamma_L$), Solow-neutral ($\gamma_L = 0$), Harrod-neutral ($\gamma_K = 0$), capital-augmenting ($\gamma_K > \gamma_L > 0$) or labor-augmenting ($\gamma_L > \gamma_K > 0$).

The key difficulty for identification is very apparent if we derive an expression for the capital-labor ratio (return to capital divided by aggregate wages, obtained from the first order condition):

$$\text{Capital-labor ratio} = \frac{\pi}{1-\pi} \left(\frac{\Gamma_{Kt} K_t}{\Gamma_{Lt} L_t} \right)^{\rho}. \tag{26}$$

An increase in the capital-labor ratio can be equally explained by a capital-augmenting technological progress if capital and labor are relative substitutes, or a labor-augmenting technological progress if they are relative complements.

In terms of the estimation, there are three possible equations that can be used (or combined) for the estimation: the two first order conditions and the production function itself. Single equation approaches concentrate either on the production function or on the ratio of the two first order conditions. In practice, assuming factor neutral technical change, when factor shares are relatively stable, tends to bias estimates towards Cobb-Douglas (Antràs, 2004), and the estimates from the labor first order condition tend to find larger estimates than those based on the capital's first order condition (León-Ledesma, McAddam, and William, 2010).

In their Monte Carlo experiments, León-Ledesma, McAddam, and William (2010) find that single equation methods does not perform very well in identifying the capital-labor elasticity in the presence of biased technical change. The worst one, in particular, is the capital first order condition, followed by the production function in levels; the labor first order condition performs better in terms of estimation biases. However, what they prove is that system methods, especially the three equation one, perform much better in practice.

D. Latent factor models

Linking the discussion in the previous two sub-sections, Krusell, Ohanian, Ríos-Rull, and Violante (2000) dig deeper in exploring what is behind the skill-biased technical change. To do that, they assume the following production technology that exhibits capital-skill complementarity:

$$Y_t = \zeta_t K_{St}^{\alpha} \left[\theta L_t^{\rho} + (1-\theta) \left(\pi H_t^{\gamma} + (1-\pi) K_{Et}^{\gamma} \right)^{\frac{\rho}{\gamma}} \right]^{\frac{1-\alpha}{\rho}}, \tag{27}$$

where K_{St} and K_{Et} are, respectively, structures and equipment capital, and L_t and H_t denote low-skilled and high-skilled labor, in efficiency units: $L_t \equiv \psi_{Lt} h_{Lt}$ and $H_t \equiv \psi_{Ht} h_{Ht}$, where h_{Lt} and h_{Ht} are total hours worked, and ψ_{Lt} and ψ_{Ht} are efficiency units. In this context, there is capital-skill complementarity provided that $\rho > \gamma$. Given that they consider a closed economy, the output identity yields:

$$Y_t = C_t + I_{St} + \frac{I_{Et}}{q_t}, \tag{28}$$

where C_t denotes aggregate consumption, I_{St} and I_{Et} denote investment in structures and equipment, and q_t are the relative prices of equipment. The skill pre-

mium is given by:

$$\ln \frac{w_{Ht}}{w_{Lt}} \approx C + (\rho - 1) \ln \frac{H_t}{L_t} + (\rho - \gamma) \frac{1 - \pi}{\pi} \left(\frac{K_{Et}}{H_t} \right)^{\gamma}, \tag{29}$$

where C is a constant, and where we used the approximation $\ln(1 + x) \approx x$. Comparing this equation to (24), we can see that what we interpreted above as skill-biased technical change, depends on the relative growth of equipment capital compared to labor, provided there is capital-skill complementarity ($\rho > \gamma$). What these authors argue is that technical change decreased the prices of equipment capital, which lead to an increase in the accumulation of equipment and capital-skill complementarity made this technological progress skill biased. In order to account for the more standard forms of skill-biased technical change, the authors specified $\ln \psi_{it} = \varphi_{0i} + \varphi_{i1} t + \epsilon_{it}$ for $i \in \{H, L\}$, where $(\epsilon_{Ht}, \epsilon_{Ht})'$ is i.i.d. bivariate normal.

They estimate the model by means of a three-equation system, as discussed above. The three equations consist of an equation for the labor share, another one for the relative wage bills for high and low skilled labor, and a third equation that equates the net rate of return of equipment and structures. Importantly, the last equation depends on the relative prices of equipment, which are an unobserved latent factor, and depreciation rates, which are additional latent factors in the model. The estimation is carried through simulated pseudo-maximum likelihood, taking into account the potential endogeneity of hours worked to technology and efficiency shocks.

E. Spatial variation

When using spatial variation, endogeneity concerns (similar to those at the firm-level) are more apparent. This is so because there are many more mechanisms of adjustment than at the national single-market level. A typical paper estimating production functions at the local level would typically ignore capital (it is rarely observed at the local level), and would consider heterogeneous labor in the following way:

$$Y_{it} = \zeta_{it} \left(\sum_j \theta_{ijt} L_{ijt}^{\rho} \right)^{\frac{1}{\rho}}, \tag{30}$$

where j typically denotes different skill groups or industries. As it is the case in any nested CES production function, the factor neutral term ζ_{it} cancels out (or it is captured by location-time fixed effects) if one works with first order conditions

(local demands). However, the terms θ_{ijt} are often considered random, as local labor market-specific shocks to different industries or skill groups. In this context, labor supply in market ij at time t is endogenous to θ_{ijt} (simultaneity bias).

A common way to deal with endogeneity is the so-called "Bartik instrument", named after Bartik (1991). In particular, consider the wage equation obtained from the first order condition on the previous production function:

$$\ln w_{ijt} = (\rho - 1) \ln L_{ijt} + \delta_{it} + \ln \theta_{ijt}, \tag{31}$$

where δ_{it} denotes time dummies that capture all variation at the market-period level. One of the many versions of the Bartik instrument for L_{ijt}, denoted by $\Delta \hat{L}_{ijt}$, is constructed as:

$$\Delta \hat{L}_{ijt} = \frac{L_{ij0}}{\sum_j L_{ij0}} \sum_{-i} \Delta L_{ijt}, \tag{32}$$

where Δ indicates over-time differences, and $\sum_{-i}$ denotes sum across all local markets excluding the market i. This instrument is often used on the estimation of (31) in first differences.

Many versions of this instrument have been used in practice. Intuitively, the instrument exploits the industrial/skill composition of market i in some (ideally far away) initial period $t = 0$ to leverage national level increases in the demand of that particular labor input. Even though this instrument is widely used in many different contexts (not only production function/labor demand estimation), it is also often criticized. Goldsmith-Pinkham, Sorkin, and Swift (2019) provide a deep discussion on when, why and how to use them.

IV. Application: Albert, Glitz, and Llull (2022)

See the paper.