

# Productivity Estimation at the Plant Level: A practical guide

Jens Matthias Arnold<sup>1</sup>  
Summer 2005

This note provides a brief overview of simple methods available for estimating plant productivity from establishment-level panel data sets. It includes examples for implementing the methods in the Stata software package.

## Introduction

Recent years have seen a surge in interest in productivity analysis at the micro level, with production establishments being the focus of attention. There are a number of econometric problems that one encounters when trying to estimate unobserved productivity as the residual of the production function, using the observed firm-level variables typically available in plant-level data sets. This note is meant to be an introduction to an admittedly arbitrary selection of these problems. The focus here is mainly on methods to remedy the so-called simultaneity and selection problems. While these are among the most prominent econometric difficulties of productivity estimation at the plant level, there are other issues not mentioned here.<sup>2</sup> Without claiming to be a comprehensive treatment of all the issues involved, this note means to offer a practical guide for estimating plant productivity consistently in the presence of the two above-mentioned biases.

## Getting Started: The Production Function

Usually, one tends to assume a functional form for the production function, in the vast majority of cases Cobb-Douglas. An alternative to the Cobb-Douglas function would be a more flexible translog function, which is in theory more

---

<sup>1</sup> Bocconi University, Milan, Italy. Contact email: jens.arnold@uni.bocconi.it. Comments and corrections are most welcome. This is a preliminary note, it makes no claim for completeness and it may contain errors.

<sup>2</sup> Another relevant problem may come from using deflated plant sales or value added as a proxy for physical output. This proxy is common use in the absence of information on physical output units, which is rarely available. In the presence of product differentiation and differences in market power among firms, using this proxy may create a bias. Correcting for these problems, however, would involve plunging deeper into the behavioural structure of the market equilibrium than what is intended here. A useful reference for this issue is Melitz (2000).

attractive because it is less restrictive. In practice, however, the restriction of the functional form to be Cobb-Douglas does not tend to make too much of a difference numerically. On the other hand, the Cobb Douglas function has the advantage that it is relatively easy to whether the estimated coefficients and the resulting returns to scale are broadly in line with common sense. For the exposition here, we assume the simplest conceivable two-factor production function of the form:

$$Y_{it} = A_{it} \cdot L_{it}^{\beta} \cdot K_{it}^{\gamma} \quad \text{where } \beta + \gamma = 1 \text{ would imply constant returns to scale.}$$

$Y_{it}$  is a measure of output like value added, while  $L_{it}$  and  $K_{it}$  represent the usage of labour and capital, respectively.  $A_{it}$  is what is called the total factor productivity (TFP) because it increases all factors' marginal product simultaneously. Using value added on the left hand side is another loss of generality. A specification where real output is regressed on labour, capital and materials avoids the assumption of additive separability of material inputs implicit in the above specification, and is thus less restrictive. Moreover, it may in certain cases be reasonable to estimate separate coefficients for labour of different skill levels, or for energy consumed in the production process. The final choice of specification is likely to be dependent on the nature of the data used.

Transforming the above production function into logarithms allows linear estimation, and henceforth small letters will be used for logs. A simple standard estimation equation of the production function then looks as follows:

$$y_{it} = \beta \cdot l_{it} + \gamma \cdot k_{it} + u_{it} \quad (1)$$

Given this equation, one can calculate an estimate for the error term  $u_{it}$ , provided the coefficients are consistently estimated. The remainder of this note will be about the problems arising when one tries to fit equation (1). The residual of this equation is the logarithm of plant-specific total factor productivity  $A_{it}$ .

### Estimating the production function: Econometric Problems

#### a) SIMULTANEITY

##### a1) What is the problem ?

The problem that is usually referred to as the simultaneity problem is that at least a part of the TFP will be observed by the firm at a point in time early enough so as to allow the firm to change the factor input decision. If that is the case, then profit maximisation of the firm implies that the realisation of the error term of the production function is expected to influence the choice of

factor inputs. This means that the regressors and the error term are correlated, which makes OLS estimates biased. Awareness of this phenomenon is far from new: It was first pointed out by Marschak and Andrews (1944).

For purposes of exposition, one can split up the error term  $u_{it}$  into two elements:

$$y_{it} = \beta \cdot l_{it} + \gamma \cdot k_{it} + \varpi_{it} + e_{it} \quad (2)$$

where  $\varpi_{it}$  is the part of the error term that is observed by the firm early enough to influence decisions, while  $e_{it}$  is a true error that may contain both unobserved shocks and measurement errors. For the econometrician,  $e_{it}$  is the noise in the signal, whereas  $\varpi_{it}$  is the systematic component that troubles his life.

a2) The remedies

#### *Fixed-effect estimation techniques*

One relatively easy way out of this problem is available if one has sufficient reason to believe that the part of TFP that influences firm behaviour,  $\varpi_{it}$ , is a plant-specific attribute, and invariant over time. In that case, including plant dummies into the regression, i.e. a fixed-effect panel regression, will solve the problem caused by  $\varpi_{it}$  and deliver consistent estimates of the parameters.

There are two drawbacks to this method: First, a substantial part of the information in the data is left unused. A fixed-effect estimator uses only the across time variation, which tends to be much lower than the cross-section one. This means that the coefficients will be weakly identified. Second, the assumption that  $\varpi_{it}$  is fixed over time may not always be reasonable, making the whole procedure invalid.

#### *The Olley and Pakes approach*

As an alternative to fixed-effect regressions, a consistent semi-parametric estimator was developed by Olley and Pakes (1996). This estimator solves the simultaneity problem by using the firm's investment decision to proxy unobserved productivity shocks. The estimation procedure involves two steps.

To begin with, we need an equation linking stocks and flows in capital. A standard approach that Olley and Pakes use for this is:

$$K_{it+1} = (1 - \delta)K_{it} + I_{it} \quad (4)$$

where  $K$  is the capital stock and  $I$  is investment. This structure means that contemporaneous values of capital and investment are orthogonal.<sup>3</sup> At the same time, the procedure assumes that expectation of future realisations of  $\varpi_{it}$  rises in its contemporaneous values. In other words, a higher value of  $\varpi_{it}$  today will induce a higher investment today even if that comes too late to affect today's capital stock. We can therefore define an (unknown) function for the optimal investment decision:

$$i_{it} = i_t(\varpi_{it}, k_{it})$$

Inverting this function, and defining  $h(\cdot) = i^{-1}(\cdot)$ , we can write  $\varpi_{it}$  as

$$\varpi_{it} = h_t(i_{it}, k_{it}) \quad (5)$$

Then the estimating equation can be rewritten as:

$$y_{it} = \beta \cdot l_{it} + \gamma \cdot k_{it} + h_t(i_{it}, k_{it}) + e_{it} \quad (6)$$

Now define the function

$$\phi(i_{it}, k_{it}) = \gamma \cdot k_{it} + h_t(i_{it}, k_{it}) \quad (7)$$

which can be approximated by 3<sup>rd</sup> or 4<sup>th</sup> order polynomials in log-labour and log-capital (including also a constant term), denoted by  $\tilde{\phi}_t$ . Hence, in the first stage one estimates the equation:

$$y_{it} = \beta \cdot l_{it} + \tilde{\phi}_t + e_{it} \quad (8)$$

Box 1

In practice, this could be done using the following Stata command lines:

```

tsset plantid year
for num 1/3: gen loginvX=loginv^X
for num 1/3: gen logcapX=logcap^X
gen cross=loginv*logcap
for num 1/3: gen crossX=cross^X

reg logY logL loginv1-loginv3 logcap1-logcap3 cross1-cross3

```

The coefficient of logarithmic labour will now be consistently estimated in equation (8). To continue, we will use the estimated function  $\tilde{\phi}$ , i.e. the

<sup>3</sup> Note that if the firm-level capital stock was constructed from investment data using the perpetual inventories method, the contemporaneous stock should contain lagged investment values only in order for the orthogonality to hold.

estimated coefficients of the investment and capital measures of all orders, to fit the values  $\tilde{\phi}_i$ . We will also need the lagged value of these fits later, and we will need to store the estimated coefficients from the last regression before we proceed.<sup>4</sup>

```

predict resid_stagel, res
gen Beta_L =_coef[logL]
gen phi=logY-resid_stagel-Beta_L*logL
gen philag1=L.phi
gen philag2=philag1^2
gen philag3=philag1^3

```

Box 2

In the second stage, it is helpful to define  $V_{it} = y_{it} - \hat{\beta} \cdot l_{it}$  and estimate the equation:

$$V_{it} = \gamma \cdot k_{it} + g(\phi_{t-1} - \gamma \cdot k_{t-1}) + \mu_{it} + e_{it} \quad (9)$$

where  $g$  is an unknown function of lagged values of  $\phi$  and capital. As in the first step, this function is again approximated by a high-order polynomial expression in  $\phi_{t-1}$  and  $k_{t-1}$ . In principle, this is the same kind of high-order polynomial approximation as the one used in the first stage. In practice, however, estimating this second stage is a bit more cumbersome than the first stage because capital appears in the regression in contemporaneous values, and is used in lagged values in the polynomial approximating the function  $g(\cdot)$ . Estimation would not be efficient if one were to ignore this known structure, i.e. if one did not restrict the coefficient on capital to be the same wherever it appears in the estimation of the second stage. This implies that we need to estimate equation (9) by non-linear least squares. Box 3 contains some advice on how to implement this using Stata.

Once equation 9 is estimated, we have estimates for all parameters of interest. We have obtained a consistent estimate for the log-labour coefficient  $\beta$  in the first stage (as well as for any additional freely adjustable production factors one may want to add) and a consistent estimate for the log-capital coefficient  $\gamma$  in the second stage. Armed with these results, we can fit equation (1) and construct the individual error terms  $u_{it}$ , which are simply the logs of our estimated plant TFP.

<sup>4</sup> Note that if there more freely adjustable factors in the production function, such as materials with plant output on the left hand side rather than value added, or different categories of labour, then the coefficient on this factor would simply be estimated in the first stage as well, and then netted out just like labour here when constructing phi.

Box 3

Stata is not quite as straightforward when it comes to implementing non-linear least squares. The main issue that will seem new to first-time users of the procedure is the fact that one needs to define starting values for the non-linear minimization procedure, define a non-linear programme and use macros (globals) to refer to the parameters within the minimization algorithm. Once one understands the logic of the non-linear least square procedure in Stata, however, the implementation of the second stage becomes much clearer. It is certainly helpful to familiarize oneself with the nl procedure at <http://www.stata.com/help.cgi?nl> before reading on.

At first, we need to create starting values for the non-linear estimation, and OLS estimates are obvious candidates for this. In other words, one should begin by running

```
gen v          = output - Beta_L*lab
gen logcaplag = L.logcap
reg v logcap philag1-philag3
```

and collect all the coefficients (including the constant) as starting values. Here, they will be referred to as `stval_` in what follows. The non-linear program needed for running the nl procedure could look like this:

```
program define nlsecstage
if "`1'" == "?" {
    global S_1 "Beta_const Beta_logcap Beta_philag1 Beta_philag2 Beta_philag3"
    global Beta_const      = stval_const
    global Beta_logcap     = stval_logcap
    global Beta_philag1    = stval_philag1
    global Beta_philag2    = stval_philag2
    global Beta_philag3    = stval_philag3
    exit
}
replace `1'=$Beta_const+ ///
        $Beta_logcap*logcap+ ///
        $Beta_philag1*(philag1-$Beta_logcap*logcaplag)+ ///
        $Beta_philag2*(philag1-$Beta_logcap*logcaplag)^2+ ///
        $Beta_philag3*(philag1-$Beta_logcap*logcaplag)^3
end
```

After the program is defined and the starting values generated, simply call it by running

```
nl secstage v if logcap!=. & philag1!=.
```

and the coefficient `Beta_K` which one can collect as `gen Beta_K = $Beta_logcap` will be a consistent estimate.

### *The Levinsohn and Petrin approach*

The method suggested by Olley and Pakes (1996) is able to generate consistent estimates for the production function estimates, provided a couple of conditions are met. One of these conditions is that there must be a strictly monotonous relationship between the proxy (investment) and output. This means that any observation with zero investment has to be dropped from the data in order for the correction to be valid. Depending on the data, this may imply a considerable drop in the number of observations because many times not every firm will have strictly positive investment in every year. Levinsohn and Petrin (2003) offer an estimation technique that is very close in spirit to the Olley and Pakes approach. Instead of investment, however, they suggest the use of intermediate inputs as a proxy rather than investment. Typically, many datasets will contain significantly less zero-observations in materials than in firm-level investment. Levinsohn and Petrin also offer several specification tests to check for the appropriateness of the proxy used. The optimal choice of proxy is very much dependent on the nature and shortcomings of the data at hand. For a more detailed discussion of the choice of proxy see Levinsohn and Petrin (2003).

Programming the Levinsohn and Petrin estimation technique in Stata is a little more cumbersome than the Olley and Pakes procedure. There is, however, a user-friendly Stata extension available at no cost, called *levpet*, which will implement production function estimations using this procedure. The command and how to download and use it is described in Levinsohn, Petrin and Poi (2003). Applying the Levinsohn and Petrin procedure with the *levpet* command is straightforward.

On occasions, the *levpet* procedure may generate strange-looking results. When using the output version of the procedure (as opposed to the value-added version), it may happen that there is not enough variation in the data for a separate identification of all coefficients. In this case, one has no choice but to change the specification, probably to a value-added form. Also, it may happen that the procedure estimates a material coefficient to be exactly one. This is due to an imposed upper limit in the estimation algorithm, and such an estimation result should be discarded as well. Unless one comes across one of these cases, however, the procedure is a good and easy way to implement a consistent estimator without much programming.

### *Generalized Method of Moments approaches*

Apart from the semi-parametric estimators described above, there are other possible avenues for estimating firm-level production functions in a consistent

manner. A GMM system estimator following Blundell and Bond (1998) is a suitable estimation method even in the presence of endogenous regressors, as may be the case under the simultaneity problem described above. It requires a longer time series, however, as sufficiently lagged values and lagged differences are used as instruments in this procedure. The degree to which these instruments are a good choice are subject to some discussion in the literature. See Blundell and Bond (2000) for an illustration of this procedure in the case of production functions.

## b) SELECTIVITY AND ATTRITION BIAS

### b1) The Problem

Many plant-level data sets contain missing values associated to plants dropping out of the sample. If these plants are selected in a non-random manner, e.g. because they stop producing, then the sample may become biased. Trying to bypass the problem by considering only a balanced sub-sample is likely to bias the estimates of the factor coefficients. This could be the case whenever plants with higher capital stock are less likely to drop out of the market (and the sample) if affected by a negative shock. In the remaining sample, there will be a non-zero (say negative) correlation between the realisations of the error term and the capital stocks. In this case, the estimated capital coefficient will suffer from a downward bias.

### b2) The remedy

Olley and Pakes' method also offers a correction for the attrition bias. This is achieved by means of a fitted value for the probability of exiting from the sample. In a first step, one estimates a probit of a survival indicator variable on a polynomial expression containing capital and investment. In a second step, the fitted values from this regression are then incorporated into equation (9) to control for the attrition bias in the second stage:

$$V_{it} = \gamma \cdot k_{it} + g(\phi_{t-1} - \gamma \cdot k_{t-1}, \hat{P}_{t-1}) + \mu_{it} + e_{it}$$

In practice, all this does is to bring a third argument into the estimation of equation (9). Hence,  $g(\cdot)$  is estimated by a high-order series expansion in  $\Phi_{t-1}, k_{t-1}, \hat{P}_{t-1}$ , including all cross terms. This procedure delivers consistently estimated coefficients of log capital, even in the presence of an attrition bias.

## Final Remarks

This short note is clearly unable to describe in due detail the methods sketched, nor is it meant to do so. It is, however, aimed to demonstrate that it is not particularly difficult to go beyond a simple OLS estimation of the production function, and gives some hands-on advice on the possible ways to go about it. Producing consistent estimates of firm-level TFP is not a tedious task with a standard econometric package such as Stata.

## References

- Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87: 115-43.
- Blundell, R. and S. Bond (2000). "GMM estimation with persistent panel data: an application to production functions" *Econometric Reviews* 19: pp.321-340
- Levinsohn, J. and Petrin, A. (2003). "Estimating Production Functions using Inputs to Control for Unobservables", *Review of Economic Studies* 70, pp.317-341
- Levinsohn, J., Petrin, A. and Poi, B. P. (2003). "Production Function Estimation in Stata using Inputs to Control for Unobservables." *Stata Journal* 4(2): pp. 113-123.
- Marschak, J. and Andrews, W. (1944). "Random Simultaneous Equations and the Theory of Production", *Econometrica* 12, pp. 143-205
- Melitz, M. (2000). "Estimating Firm-Level Productivity in Differentiated Product Industries." Mimeo, Harvard University.
- Olley, S. and Pakes, A. (1996). "The Dynamics Of Productivity In The Telecommunications Equipment Industry", *Econometrica* 64, pp.1263-1297