Production Function Estimation in Stata Using Inputs to Control For Unobservables

Amil Petrin
University of Chicago
National Bureau of Economic Research

James Levinsohn
University of Michigan
National Bureau of Economic Research

Brian P. Poi
Stata Corporation

November 13, 2003

Abstract

A key issue in the estimation of production functions is the correlation between unobservable productivity shocks and input levels. Profit-maximizing firms respond to positive productivity shocks by expanding output, which requires additional inputs. Negative shocks lead firms to pare back output, decreasing their input usage. Olley and Pakes (1996) develop an estimator that uses investment as a proxy for these unobservable shocks. More recently, Levinsohn and Petrin (2003) introduce an estimator which uses intermediate inputs as proxies, arguing that intermediates may respond more smoothly to productivity shocks. This paper reviews Levinsohn and Petrin's approach and introduces a Stata command that implements it.

1 Introduction

A key issue in the estimation of production functions is the correlation between unobservable productivity shocks and input levels.[1] Profit-maximizing firms respond to positive productivity shocks by expanding output, which requires additional inputs. Negative shocks lead firms to pare back output, decreasing their input usage. When true, Ordinary Least Squares (OLS) estimates of production functions are biased, and, by implication, lead to biased estimates of productivity, often the relevant quantity for the estimation question.

Olley and Pakes (1996) (OP) develop an estimator that uses investment as a proxy for these unobservable shocks. More recently, Levinsohn and Petrin (2003) (LP) point to the evidence from firm-level datasets that suggest investment is very lumpy, that is, that there are substantial adjustment costs. If true, the investment proxy may not smoothly respond to the productivity shock, violating the consistency condition.

LP show the conditions under which intermediate inputs can also solve this simultaneity problem. Remarkably, in most applications these inputs are not used beyond subtracting them from the gross output number to get value added, so the approach comes at no additional cost in terms of data or computation. LP discuss the theoretical benefits of extending the proxy choice set in this direction, and provide substantial empirical evidence that these benefits are important.

One benefit is strictly data-driven. It turns out that the investment proxy is only valid for plants reporting non-zero investment. (This is due to an invertibility condition described below.) Pronounced adjustment costs, which do not invalidate the use of investment as a proxy, are the likely reason that over one-half of the chilean sample used later reports zero investment. This kind of severe truncation is not unique to the Chilean data. Much of the plant-level research being conducted today is on data from countries like India, Turkey, Columbia, Mexico, and Indonesia, and in these data sets - as is likely with others - the "zero investment" problem looms large.

Using intermediate input proxies instead of investment avoids truncating all of the zero investment firms. In the data above (at least), firms almost always report positive use of intermediate inputs like electricity or materials.

To the extent that adjustment costs are an important issue, intermediate inputs may confer another benefit. If it is less costly to adjust the intermediate input, it may respond more fully to the entire productivity term than investment. For example, if adjustment costs lead to kink points in the investment demand function, plants may not respond fully to productivity shocks, and some correlation between the regressors and the error term can

[1] For an overview of the history of this discussion, see Griliches and Mareisse (1998).

remain.

Another nice feature of the intermediate input is it provides a simple link between the estimation strategy and the economic theory, primarily because intermediate inputs are not typically state variables. Levinsohn and Petrin (2003) develop this link, showing the (mild) conditions that must hold if intermediate inputs are to be a valid proxy for the productivity shock. They suggest three specification tests for evaluating any proxy's performance. In addition, they derive the expected directions of bias on the OLS estimates relative to LP's intermediate input approach when simultaneity exists. Finally, LP show for the four largest Chilean manufacturing industries that significant differences between OLS and Levinsohn-Petrin exist that are exactly consistent with simultaneity.[2]

This paper reviews Levinsohn and Petrin's approach and provides a Stata command that implements it.

2 Productivity estimation

In this section we give an overview with an emphasis on the mechanics of the estimator. A more detailed exposition can be found in Levinsohn and Petrin (2003).

For the purposes of this note the production technology is assumed to be Cobb-Douglas:

$$y_t = \beta_0 + \beta_l l_t + \beta_k k_t + \beta_m m_t + \omega_t + \eta_t,$$

where y_t is the logarithm of the firm's output, most often measured as gross revenue or value added, l_t and m_t are the logarithm of the freely variable inputs labor and the intermediate input, and k_t is the logarithm of the state variable capital.[3] For ease of exposition we include only two freely variable inputs, though the command accompanying this article allows for an arbitrary number of them.

The error has two components, the transmitted productivity component given as ω_t, and η_t, an error term that is uncorrelated with input choices. The key difference between ω_t and η_t is that the former is a state variable and hence impacts the firm's decision rules. It is not observed by the econometrician, and it can impact the choices of inputs, which leads to the well-known simultaneity problem in production function estimation. Estimators ignoring this correlation between inputs and this unobservable factor (like OLS) will yield inconsistent results.

[2]This finding has been reported for a number of other manufacturing surveys, including ones from Columbia and India.

[3]The approach extends immediately to other forms of the production technology. For example, with an appropriate definition of variables, trans-log (and higher order) production functions can be estimated with this stata routine.

Demand for the intermediate input m_t is assumed to depend on the firm's state variables k_t and ω_t:

$$m_t = m_t(k_t, \omega_t)$$

Making mild assumptions about the firm's production technology, Levinsohn and Petrin (2003, Appendix A) show that the demand function is monotonically increasing in ω_t. This allows inversion of the intermediate demand function, so ω_t can be written as a function of k_t and m_t:

$$\omega_t = \omega_t(k_t, m_t).$$

The unobservable productivity term is now expressed solely as a function of two observed inputs.

A final identification restriction follows Olley-Pakes. LP assume that productivity is governed by a first-order Markov process:

$$\omega_t = \mathrm{E}[\omega_t | \omega_{t-1}] + \xi_t$$

where ξ_t is an innovation to productivity that is uncorrelated with k_t (but not necessarily with l_t; this is part of the source of the simultaneity problem).

First we discuss estimation when the dependent variable is value added. Then we turn to using output (or gross revenue) as the dependent variable.

2.1 Estimation in the value-added case

Letting v_t represent value added - gross output net of intermediate inputs - we can write the production function as

$$\begin{aligned} v_t &= \beta_0 + \beta_l l_t + \beta_k k_t + \omega_t + \eta_t \\ &= \beta_l l_t + \phi_t(k_t, m_t) + \eta_t \end{aligned}$$

where

$$\phi_t(k_t, m_t) = \beta_0 + \beta_k k_t + \omega_t(k_t, m_t).$$

By substituting a third-order polynomial approximation in k_t and m_t in place of $\phi_t(k_t, m_t)$, it is possible to consistently estimate parameters of the value added equation using OLS as

$$v_t = \delta_0 + \beta_l l_t + \sum_{i=0}^{3} \sum_{j=0}^{3-i} \delta_{ij} k_t^i m_t^j + \eta_t,$$

where β_0 is not separately identified from the intercept of $\phi_t(k_t, m_t)$.[4] This completes the first stage of the estimation routine from Levinsohn-Petrin (2003), from which an estimate of β_l and an estimate of ϕ_t (up to the intercept) are available.

The second stage of the routine identifies the coefficient β_k. It begins by computing the estimated value for ϕ_t using

$$\begin{aligned}\widehat{\phi_t} &= \widehat{v_t} - \widehat{\beta_l} l_t \\ &= \widehat{\delta_0} + \sum_{i=0}^{3}\sum_{j=0}^{3-i} \widehat{\delta_{ij}} k_t^i m_t^j - \widehat{\beta_l} l_t.\end{aligned}$$

For any candidate value β_k^*, one can compute (up to a scalar constant) a prediction for ω_t for all periods t using

$$\widehat{\omega}_t = \widehat{\phi}_t - \beta_k^* k_t.$$

Using these values, a consistent (non-parametric) approximation to $\mathrm{E}[\omega_t|\omega_{t-1}]$ is given by the predicted values from the regression

$$\widehat{\omega}_t = \gamma_0 + \gamma_1\omega_{t-1} + \gamma_2\omega_{t-1}^2 + \gamma_3\omega_{t-1}^3 + \epsilon_t,$$

which LP call $\mathrm{E}[\widehat{\omega_t|\omega_{t-1}}]$.

Given $\widehat{\beta}_l$, β_k^*, and $\mathrm{E}[\widehat{\omega_t|\omega_{t-1}}]$, LP write the sample residual of the production function as

$$\widehat{\eta_t + \xi_t} = v_t - \widehat{\beta}_l l_t - \beta_k^* k_t - \mathrm{E}[\widehat{\omega_t|\omega_{t-1}}].$$

Our estimate $\widehat{\beta}_k$ of β_k is defined as the solution to

$$\min_{\beta_k^*} \sum_t (v_t - \widehat{\beta}_l l_t - \beta_k^* k_t - \mathrm{E}[\widehat{\omega_t|\omega_{t-1}}])^2.$$

The Stata command accompanying this note uses a golden section search algorithm to minimize that function. A bootstrap approach (discussed shortly) is used to construct standard errors for $\widehat{\beta}_l$ and $\widehat{\beta}_k$. We now turn to point estimation when the dependent variable is output (or gross revenue).

2.2 Estimation in the gross revenue case

Letting y_t denote revenue, the production function is given as

$$\begin{aligned}y_t &= \beta_0 + \beta_l l_t + \beta_k k_t + \beta_m m_t + \omega_t + \eta_t \\ &= \beta_l l_t + \phi_t(k_t, m_t) + \eta_t\end{aligned}$$

[4] Another restriction is necessary to separately identify β_0 from the intercept of $\phi_t(k_t, m_t)$.

where now

$$\phi_t(k_t, m_t) = \beta_0 + \beta_k k_t + \beta_m m_t + \omega_t(k_t, m_t).$$

Estimation of $\widehat{\beta}_l$ proceeds exactly as before, using OLS with a third-order polynomial approximation in k_t and m_t in place of $\phi_t(k_t, m_t)$.

The first part of the second stage is also similar to the value added case. For any candidate values β_k^* and β_m^* (for β_k and β_m), estimate $\widehat{\omega}_t$ using

$$\widehat{\omega}_t = \widehat{\phi}_t - \beta_k^* k_t - \beta_m^* m_t.$$

Using the ω_t's for all t, estimate $\mathrm{E}[\widehat{\omega_t|\omega_{t-1}}]$ as before. Then, the residual for (β_k^*,β_m^*) is computed as

$$\widehat{\eta_t + \xi_t} = y_t - \widehat{\beta}_l l_t - \beta_k^* k_t - \beta_m^* m_t - \mathrm{E}[\widehat{\omega_t|\omega_{t-1}}].$$

This residual must be interacted with at least two instruments to identify both β_k and β_m. Similar to the value added case, if period t's capital stock is determined by the previous period's investment decisions, it does not respond to shocks to this period's productivity innovation term ξ_t, providing the moment condition

$$\mathrm{E}[\eta_t + \xi_t | k_t] = 0,$$

which is implicitly imposed in the objective function from (2.1). An additional moment condition is needed to identify β_m separately from β_k. LP use the fact that last period's level of material usage m_t is uncorrelated with this period's error, giving us the moment condition

$$\mathrm{E}[\eta_t + \xi_t | m_{t-1}] = 0.$$

Thus, with $\boldsymbol{Z}_t \equiv (k_t, m_{t-1})$, one candidate estimator solves

$$\min_{(\beta_k^*,\beta_m^*)} \sum_h \left[\sum_t (\widehat{\eta_t + \xi_t}) Z_{ht}\right]^2,$$

with h indexing the elements of Z_t.

Additional overidentification conditions are given by

$$\mathrm{E}[\eta_t + \xi_t | l_{t-1}] = 0, \quad \mathrm{E}[\eta_t + \xi_t | m_{t-2}] = 0, \quad \text{and} \quad \mathrm{E}[\eta_t + \xi_t | k_{t-1}] = 0.$$

These can be used to improve efficiency and test the specification. Here one redefines $\boldsymbol{Z}_t \equiv (k_t, m_{t-1}, l_{t-1}, m_{t-2}, k_{t-1})$. $\widehat{\beta}_k$ and $\widehat{\beta_m}$ are then defined as the solution to

$$\min_{(\beta_k^*,\beta_m^*)} \sum_h \left[\sum_t (\widehat{\eta_t + \xi_t}) Z_{ht}\right]^2.$$

Our Stata implementation provides two methods for solving the GMM minimization problem. The default behavior is to use Stata's `nl` command, which is based on Newton's method. Alternatively, a two-dimensional grid search can be requested. Candidate values for β_k^* and β_m^* from 0.01 to 0.99 in increments of 0.01 are used. Although much slower than `nl`, the grid search is handy for confirming that `nl` has found the global minimum of the objective function. Moreover, if there is insufficient variation in the capital and proxy veriables, then `nl` may have difficulty solving the minimization problem; in these cases one can instead use the grid search.

2.3 Standard Errors

The estimators developed above involve two main stages of estimation. In each of these stages a number of preliminary estimators are used. The covariance matrix of the final parameters must account for the sampling variation introduced by all of the estimators used in the two stages. Although deriving an analytic covariance matrix may be feasible, this calculation is not trivial. Instead, LP substitute computational power for analytic difficulties, employing the bootstrap to estimate standard errors.

Since LP use panel data, they sample with replacement from firms, using the entire time series of observations for that firm in the bootstrapped sample when the firm's id number is randomly drawn. A bootstrapped sample is complete when the number of firm-year observations equals (or closely equals) the number of firm-year observations in the original sample. The variation in the point estimates across the bootstrapped samples provides an estimate for the standard errors of the original point estimates. In Stata this is accomplished using the `cluster(`*varname*`)` option with the `bootstrap` or `bsample` commands.

Bootstrapping in the case where overidentifying restrictions are imposed is slightly different. The sample moments computed using the original dataset will in general not equal zero even though the population moments do (by assumption).[5] As Horowitz (2001) and others have noted, this means that, for each of the bootstrapped samples, one must "re-center" the moment conditions by subtracting the values of the sample moments calculated using the original dataset (at the minimum). Our Stata implementation accomplishes this by first performing the estimation on the original dataset. We then store the values of the sample moments (using a series of global macros). Their value is then subtracted from the bootstrapped sample's moments when minimizing the objective function for that bootstrapped sample. This restores the consistency of the bootstrap approach in the construction of standard errors.

[5] See Horowitz (2001) for an overview of the bootstrap and a discussion of the necessity of recentering.

3 Stata implementation

3.1 Syntax

`levpet` *depvar* [`if`] [`in`] , `free(`*varlist*`)` `proxy(`*varlist*`)` `capital(`*varname*`)` [[`valueadded` | `revenue`] `justid` `grid` `i(`*varname*`)` `t(`*varname*`)` `reps(`#`)` `level(`#`)`]

Syntax for predict

`predict` [`type`] *newvarname* [`if`] [`in`] , `omega`

3.2 Options

`free(`*varlist*`)` specifies the freely variable inputs, excluding the one used as the proxy variable.

`proxy(`*varlist*`)` specifies the proxy variable, typically electricity, materials, or fuels.

`capital(`*varname*`)` specifies the capital variable.

`justid` requests that the GMM estimator use only present-period capital and the first lag of the proxy variable be used as instruments. The default is to include lagged labor, lagged capital, and the second lag of the proxy variable as instruments as well. This option can only be used with `revenue`.

`grid` requests that the GMM estimator use a grid search to minimize the criterion function with respect to the coefficients on capital and the intermediate input. The default is to use Stata's `nl` command. This option can only be used with `revenue`.

`i(`*varname*`)` specifies the variable that contains the unit to which the observation belongs. You can specify the `i()` the first time you estimate, or you can use the `iis` command to set `i()` beforehand. Note that it is not necessary to specify `i()` if the data has been previously `tsset`, or if `iis` has been previously specified—in these cases, the group variable is taken from the previous setting. See [XT] **xt**.

`t(`*varname*`)` specifies the variable that contains the time at which the observation was made. You can specify the `t()` the first time you estimate, or you can use the `tis` command to set `t()` beforehand. Note that it is not necessary to specify `t()` if the data has been previously `tsset`, or if `tis` has been previously specified—in these cases, the time variable is taken from the previous setting. See [XT] **xt**.

`reps(`#`)` specifies the number of bootstrap replications to be performed. The default is 50.

`valueadded` indicates that the dependent variable represents value added and that the least-squares estimator be used. This is the default.

`revenue` indicates that the dependent variable represents gross revenue and that the GMM estimator be used.

`level(`#`)` specifies the confidence level, in percent, for confidence intervals. The default is `level(95)` or as set by `set level`; see [U] **23.6 Specifying the width of confidence intervals**.

Options for predict

`omega` requests the predicted levels of productivity, where for the value added case

$$\widehat{\omega_t} = \exp(v_t - \widehat{\beta_l} l_t - \widehat{\beta_k} k_t)$$

and for the gross revenue case

$$\widehat{\omega_t} = \exp(y_t - \widehat{\beta_l} l_t - \widehat{\beta_k} k_t - \widehat{\beta_m} m_t)$$

Predict assumes the production function inputs are in log-levels, and adjusts ω_t accordingly. If there is more than one freely variable input, these formulæare modified accordingly.

3.3 Remarks

The `levpet` command implements the Levinsohn-Petrin estimator as discussed in the previous section. The command works with versions 7.0 and higher of Stata. The dialog box requires Stata 8 and can be invoked by typing in

```
. db levpet
```

4 Example

Here we illustrate the usage of `levpet` using a dataset consisting of Chilean apparel firms from 1987 through 1996. We have data on value added as well as firms' usage levels of blue and white collar labor, electricity, and capital. Here we treat blue and white collar labor as freely variable inputs and we use electricity as the proxy variable. In the examples below, the prefix `ln` on a variable name indicates the natural log.

We consider the case where the dependent variable represents value added. In Stata, we type in

```
. tsset ppn year
. levpet lnva, free(lnb lnw) proxy(lne) capital(lnk) /*
    */ valueadded reps(250)
```

The output looks very similar to most of Stata's [XT] commands:

```
Levinsohn-Petrin productivity estimator

Dependent variable represents value added.      Number of obs      =      2713
Group variable (i): ppn                         Number of groups   =       556
Time variable (t): year
                                                Obs per group: min =         1
                                                               avg =       4.9
                                                               max =        10

------------------------------------------------------------------------------
        lnva |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnb |   .4659176   .0443553    10.50   0.000     .3789828    .5528524
         lnw |    .420271   .0345444    12.17   0.000     .3525653    .4879768
         lnk |   .2250087   .0646426     3.48   0.000     .0983115    .3517059
------------------------------------------------------------------------------
Wald test of constant returns to scale: Chi2 =   2.96 (p = 0.0856).
```

The header of the output summarizes the panel data structure of the dataset, and below that are the estimated parameters. At the bottom of the output is a Wald test of constant returns to scale; it is simply a test that the sum of the coefficients equals one.

In Table 1 we compare parameter estimates from OLS, fixed-effects regression, and the LP estimator. For the parameters on the freely variable inputs, the OLS estimates exceed the LP estimates, confirming both the theoretical and empirical results discussed in Levinsohn and Petrin (2003). Whether the OLS coefficient on capital will be biased upward or downward depends on the degree of correlation among the inputs and the productivity shocks. In this particular application, the OLS estimate is less than the LP estimate. The fixed-effects estimates differ quite substantially from both the OLS and LP estimates. One explanation is that the magnitude of each firm's productivity shock varies over time and is not a constant fixed effect.

At the bottom of Table 1 we also report the sum of the coefficients for each estimator; constant returns to scale corresponds to a sum of one. Both OLS and LP imply increasing returns for this industry, though in the case of LP we cannot reject the null that the sum is one at the 5% significance level.

5 Conclusion

When estimating production functions, one must account for the correlation between input levels and productivity. Profit-maximizing firms respond to increases in productivity by increasing their usage of factor inputs. Methods that ignore this endogeneity, such as OLS and the fixed-effects estimator, will provide inconsistent estimates of the parameters of the production function.

Building on the work of Olley and Pakes (1996), Levinsohn and Petrin (2003) develop an estimator that utilizes intermediate inputs to proxy for the unobservable productivity term. Most plant-level datasets include data on the usage of intermediate inputs such as energy and materials, so Levinsohn and Petrin's estimator does not suffer from the truncation bias induced by Olley and Pakes' estimator, which requires firms to have non-zero levels of investment.

In this paper we have introduced the Stata command `levpet` to implement this estimator. We hope that its simple syntax will motivate people to consider it as a better alternative to estimators that ignore endogeneity issues.

6 References

Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman and E. Leamer (Eds.). *Handbook of Econometrics,* Volume 5, Amsterdam: North-Holland.

Griliches, Z., & Mareisse, J. (1998). Production functions: The search for Identification. *In Econometrics and economic theory in the twentieth century: The Ragnar Prisch Centennial Symposium,* pp. 169-203. Cambridge University Press.

Levinsohn, J., & Petrin, A. (2003a). Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies,* Vol. 70(2), No. 243, pp. 317-342.

Levinsohn, J., & Petrin, A. (2003b). On the Micro-Foundations of Productivity Growth. *University of Chicago Mimeo.*

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

Appendix — Saved results

`levpet` saves in `e()`:

Scalars

`e(N)`	number of observations
`e(waldcrs)`	Wald test of constant returns

Macros

`e(cmd)`	`levpet`
`e(predict)`	`levpet_p`
`e(model)`	`value added` or `revenue`; model used

Matrices

`e(b)`	coefficient vector
`e(V)`	variance-covariance matrix

Table 1
Comparison of OLS, Fixed-Effects, and LP Estimators
Dependent Variable is Log of Value Added
(Standard Errors in Parentheses)

Parameter	Model		
	OLS	FE	LP
`lnb`	0.5612	0.4989	0.4659
	(0.0191)	(0.0275)	(0.0443)
`lnw`	0.4895	0.2423	0.4202
	(0.0181)	(0.0262)	(0.0345)
`lnk`	0.1743	0.1015	0.2250
	(0.0122)	(0.0208)	(0.0646)
Sum	1.2251	0.8427	1.1111
	(0.0139)	(0.0399)	(0.0646)