The Stata Journal (2008)
8, Number 2, pp. 221–231

Production function estimation in Stata using the Olley and Pakes method

Mahmut Yasar
Department of Economics
University of Texas at Arlington
Arlington, TX
myasar@uta.edu

Rafal Raciborski
Department of Political Science
Emory University
Atlanta, GA

Brian Poi
StataCorp
College Station, TX

Abstract. Productivity is often computed by approximating the weighted sum of the inputs from the estimation of the Cobb–Douglas production function. Such estimates, however, may suffer from simultaneity and selection biases. Olley and Pakes (1996, *Econometrica* 64: 1263–1297) introduced a semiparametric method that allows us to estimate the production function parameters consistently and thus obtain reliable productivity measures by controlling for such biases. This study first reviews this method and then introduces a Stata command to implement it. We show that when simultaneity and selection biases are not controlled for, the coefficients for the variable inputs are biased upward and the coefficients for the fixed inputs are biased downward.

Keywords: st0145, opreg, levpet, production function, bias, simultaneity

1 Introduction

Productivity is often estimated as the deviation between observed output and output predicted by a Cobb–Douglas production function estimated by ordinary least squares (OLS). Such estimates, however, may suffer from simultaneity and selection biases. Olley and Pakes (1996) introduced a semiparametric method that controls for these biases, allowing us to estimate the production function parameters consistently and thus obtain reliable productivity estimates.

Simultaneity arises because productivity is known to the profit-maximizing firms (but not to the econometrician) when they choose their input levels (Marschak and Andrews 1944). Firms will increase their use of inputs as a result of positive productivity shocks. OLS estimation of production functions will yield biased parameter estimates because it does not account for the unobserved productivity shocks. A fixed-effect estimator would solve the simultaneity problem only if we are willing to assume that the unobserved, firm-specific productivity is time-invariant. Other methods, including instrumental-variables approaches, have also been proposed to control for this bias when estimating the parameters of production functions.[1]

Another issue that one needs to address when estimating production function parameters is selection bias. Selection bias results from the relationship between productivity shocks and the probability of exit from the market. If a firm's profitability is positively related to its capital stock, then a firm with a larger capital stock is more likely to stay in the market despite a low productivity shock than a firm with a smaller capital stock, because the firm with more capital can be expected to produce greater future profits. The negative correlation between capital stock and probability of exit for a given productivity shock will cause the coefficient on the capital variable to be biased downward unless we control for this effect.

Olley and Pakes (1996) proposed a novel approach to address the simultaneity and selection problems while estimating the production function parameters and firm-level productivity.[2] The simultaneity problems are addressed by using investment to proxy for an unobserved time-varying productivity shock, and the selection problems are addressed by using survival probabilities.

This paper first reviews this methodology and then introduces a Stata command to implement it.

2 Estimation

The Olley and Pakes (1996) approach assumes that incumbent firms decide at the beginning of each period whether to continue participating in the market. If the firm exits, it receives a liquidation value of Φ dollars and never appears again. If it does not exit, it chooses variable inputs (such as labor, material, and energy) and a level of investment, I_{it}. The firm also realizes profits conditional on the beginning-of-period state variables: a productivity indicator or shock, Ω_{it}; the capital stock, K_{it}; and the age of the firm, a_{it}. We further assume that expected productivity is a function of current productivity and capital, $E[\Omega_{i,t+1}\,|\,\Omega_{it}, K_{it}]$, and that the firm's profit is a function of Ω_{it} and K_{it}.

Firm i's decision to maximize the expected discounted value of net future profits is then characterized by the Bellman equation:

[1] See Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998, 2000), Griliches and Mareisse (1998), Levinsohn and Petrin (2003), Pavcnik (2002), and Wooldridge (2005) for further information.

[2] Levinsohn and Petrin (2003) proposed a similar approach, which uses intermediate inputs instead of investment to control for correlation between inputs and the unobserved productivity shock, and thus limits the problems associated with lumpy investment. Petrin, Poi, and Levinsohn (2004) introduced a Stata program, `levpet`, to implement the methodology proposed by Levinsohn and Petrin (2003) that controls for simultaneity bias (but not for selection bias).

$$V_{it}(K_{it}, a_{it}, \Omega_{it}) = \\ \text{Max}[\Phi, \text{Sup}_{I_{it} \geq 0} \Pi_{it}(K_{it}, a_{it}, \Omega_{it}) - C(I_{it}) + \rho E\{V_{i,t+1}(K_{i,t+1}, a_{i,t+1}, \Omega_{i,t+1}) \,|\, J_{it}\}]$$

where $\Pi_{it}(\cdot)$ is the profit function (current profits as a function of the state variables), $C(\cdot)$ is the cost of current investment, ρ is the discount factor, and $E[\,\cdot\,|\,J_{it}]$ is the firm's expectations operator conditional on information J_{it} at time t. The Bellman equation implies that a firm exits the market if its liquidation value, Φ, exceeds its expected discounted returns. The solution to this equation results in a Markov perfect equilibrium strategy defining rules for exit and for investment decisions.

Specifically, firm i will decide to stay in the market ($\chi_{it} = 1$) or exit the market ($\chi_{it} = 0$) if its productivity is greater than or less than some threshold subject to the firm's current capital stock and age, K_{it} and a_{it}. This exit rule is written as follows:

$$\chi_{it} = \left\{ \begin{array}{ll} 1 & \text{if } \Omega_{it} \geq \underline{\Omega}_{it}(K_{it}, a_{it}) \\ 0 & \text{otherwise} \end{array} \right\} \tag{1}$$

We assume the state variable Ω_{it} follows a first-order Markov process.

The firm's decision to invest in further capital, I_{it}, depends on Ω_{it}, K_{it}, and a_{it}:

$$I_{it} = I(\Omega_{it}, K_{it}, a_{it}) \tag{2}$$

This investment decision equation implies that future productivity is increasing in the current productivity shock, so firms that experience a large positive productivity shock in period t will invest more in period $t + 1$.

Based on these exit and investment decision rules, Olley and Pakes (1996) specify a production function (OP) to estimate the parameters consistently. Assume that the production technology is represented by a production function that relates output to inputs and the productivity residual or shock:

$$Y_{it} = F(L_{it}, M_{it}, E_{it}, K_{it}, a_{it}, \Omega_{it})$$

For estimation purposes, we assume Cobb–Douglas technology

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_m m_{it} + \beta_e e_{it} + \beta_k k_{it} + \beta_a a_{it} + u_{it} \tag{3}$$

$$u_{it} = \Omega_{it} + \eta_{it} \tag{4}$$

where y_{it} is log output for firm i in period t; l_{it}, m_{it}, e_{it}, and k_{it} are the log values of labor, material, energy, and capital inputs; a_{it} is the age of the firm; Ω_{it} is the productivity shock that is observed by the decision-maker in the firm but not by the econometrician; and η_{it} is an unexpected productivity shock that is unobserved by both the decision-maker and the econometrician. Thus η_{it} has no effect on the firm's decisions, but Ω_{it} is a state variable that does affect the firm's decision-making process.

Given the assumptions of the model, standard econometric models provide biased and inconsistent estimates of (3) for two reasons: simultaneity between output and variable inputs, and selection bias resulting from the exit of inefficient firms. Specifically, the assumption that Ω_{it} is seen by the firm but not by the econometrician implies that inputs are correlated with the realization of the productivity shock (this argument was first formalized by Marschak and Andrews [1944]). If the firms' higher variable input use resulting from a positive productivity shock (Ω_{it}) is not accounted for in the production function, the OLS estimates for these inputs will be biased upward because of this simultaneity issue. In addition, if profitability is positively related to K_{it}, then a firm with a higher capital stock will expect larger future profitability at current productivity levels, and thus will survive lower productivity realizations that cause small firms to exit the market. This selection effect will cause expected future productivity to be negatively related to K_{it} and, thus, the capital coefficient to be biased downward.

Unlike standard estimation methods such as OLS, OP accounts for these issues. Applying this method first involves using the investment decision rule, (2), to control for the correlation between the error term and the inputs. This is based on the assumption that future productivity is strictly increasing with respect to Ω_{it}, so firms that observe a positive productivity shock in period t will invest more in that period, for any K_{it} and a_{it}. Provided that I_{it} is strictly positive, we can write the inverse function for the unobserved shock, Ω_{it}, as

$$\Omega_{it} = I^{-1}(I_{it}, K_{it}, a_{it}) = h(I_{it}, K_{it}, a_{it}) \tag{5}$$

which is strictly increasing in I_{it}.

This function can thus be used to control for the simultaneity problem. Substituting (4) and (5) into (3) yields

$$y_{it} = \beta_l l_{it} + \beta_m m_{it} + \beta_e e_{it} + \phi(i_{it}, k_{it}, a_{it}) + \eta_{it} \tag{6}$$

where $\phi(i_{it}, k_{it}, a_{it}) = \beta_0 + \beta_k k_{it} + \beta_a a_{it} + h(i_{it}, k_{it}, a_{it})$, and we approximate $\phi(\cdot)$ with a second-order polynomial series in age, capital, and investment. The partially linear (6) can be estimated by OLS. The coefficient estimates for variable inputs (labor, material, and energy) will be consistent because $\phi(\cdot)$ controls for unobserved productivity, and thus the error term is no longer correlated with the inputs.

Equation (6) does not identify β_k and β_a, so more work is required to disentangle the effects of capital and age on the investment decision from their effect on output. Achieving this requires a second step to estimate survival probabilities, which will then allow us to control for selection bias. Recall the exit rule, (1), which implies that a firm will choose to stay in the market if its productivity is greater than some threshold, $\underline{\Omega}_{it}$, that depends on K_{it} and a_{it}. The probability of survival in period t thus depends on $\Omega_{i,t-1}$ and $\underline{\Omega}_{i,t-1}$, and in turn on age, capital, and investment at time $t-1$. In our implementation, we estimate the probability of survival by fitting a probit model of χ_{it} on $I_{i,t-1}$, $K_{i,t-1}$, and $a_{i,t-1}$, as well as on their squares and cross products.[3] Call the predicted probabilities from this model $\widehat{P}_{it}$.

In the third step, we fit the following equation by nonlinear least squares:

$$\begin{aligned} y_{it} - \widehat{\beta}_l l_{it} - \widehat{\beta}_m m_{it} - \widehat{\beta}_e e_{it} = &\\ \beta_k k_{it} + \beta_a a_{it} + g(\widehat{\phi}_{t-1} - \beta_k k_{i,t-1} - \beta_a a_{i,t-1}, \widehat{P}_{it}) + \xi_{it} + \eta_{it} & \qquad (7) \end{aligned}$$

where the unknown function $g(\cdot)$ is approximated by a second-order polynomial in $\widehat{\phi}_{t-1} - \beta_k k_{i,t-1} - \beta_a a_{i,t-1}$ and $\widehat{P}_{it}$.[4] The function $g(\cdot)$ is similar in spirit to the inverse of Mills' ratio that is included in two-step sample selection models, but it is complicated by the fact that here the sample selection bias depends on two unknown variables (Ω_{it} and $\underline{\Omega}_{it}$) rather than on just one (the probability of being in the selected subsample).

Because the estimation routine involves three steps, deriving the appropriate analytic standard errors is nontrivial.[5] Our command, therefore, uses the clustered bootstrap, treating all observations for a single firm as one cluster. Fitting (7) tends to be somewhat slow with large datasets, so patience is required.

3 Stata Implementation

3.1 Syntax

```
opreg depvar [if] [in], exit(varname) state(varlist) proxy(varname)
   free(varlist) [cvars(varlist) vce(bootstrap, bootstrap_options) level(#)]
```

3.2 Options

`exit(`*varname*`)` specifies a dummy variable indicating whether firm i exited from the market in year t. A value of 1 indicates the firm exited.

`state(`*varlist*`)` specifies the state variables that appear in the production function. Typical state variables are age and the log of capital.

[3] One can, alternatively, use a kernel estimator for the second stage (see page 1278 in Olley and Pakes [1996] for an explanation).

[4] One can, alternatively, use a kernel estimator for the third stage (see page 1279 in Olley and Pakes [1996] for a discussion).

[5] Wooldridge (2005) shows how to obtain standard errors for the two-step Levinsohn and Petrin (2003) method based on the generalized method of moments. A similar argument could be used here, though implementing such an estimator in Stata would be challenging.

`proxy(`*varname*`)` specifies the variable that proxies for unobserved productivity. Typically this variable is the log of investment.

`free(`*varlist*`)` specifies the freely variable inputs, such as the logs of materials, energy, and labor.

`cvars(`*varlist*`)` specifies any additional independent variables that will be used in the first and second stages of estimation. Examples include year, size of firm, and region dummy variables.

`vce(bootstrap,` *bootstrap_options*`)` allows specification of options to control the bootstrap process. The most commonly used *bootstrap_option* is `reps(`*#*`)`, which controls the number of replications performed; the default is `reps(50)`.

`level(`*#*`)` sets the confidence level; the default is `level(95)`. See [R] **estimation options**.

3.3 Description

`opreg` implements the production function estimator of OP. The command works with Stata versions 9.2 and higher. A panel variable and a time variable must be specified. Use `xtset` to declare a variable as panel data; see [XT] **xtset**.

3.4 Remarks

Our implementation approximates the unknown functions $\phi(\cdot)$ and $g(\cdot)$ by using polynomial expansions. If two state variables x and y are specified along with the proxy variable z, then we use

$$\phi(x,y,z) \approx c_0 + c_1x + c_2y + c_3z + c_4x^2 + c_5y^2 + c_6z^2 + c_7xy + c_8xz + c_9yz$$

where the c's are parameters estimated along with the other model parameters. When a single state variable x is specified, we use

$$\phi(x,z) \approx c_0 + c_1x + c_3z + c_4x^2 + c_6z^2 + c_8xz$$

4 Example

We illustrate our command with an unbalanced panel of firms in the COMPUSTAT North America database during the period 1995–2002. The dataset consists of 3,772 firms and 19,710 observations for which output is available. We divided sales and nominal values of inputs by their corresponding deflators to obtain constant-dollar quantities. The deflators are taken from the Bartelsman and Gray (2001) productivity database. In this sample, nearly 98% of the observations include nonzero levels of investment, so we suspect that sample selection biases will be modest. Here we show how to use our command. For those who are interested in the details of the computations, turn to the appendix; there we show the steps `opreg` performs behind the scenes.

In our dataset, the variable `lny` represents log output; `exit` is a dummy variable with 1 indicating the firm exited in the current period and 0 otherwise; `t` is the trend; `age` is the age of the firm; and `lnkop`, `lnm`, `lnl`, and `lninv` are the logs of capital, materials, labor, and investment, respectively. We treat `age` and `lnkop` as state variables, `lnl` and `lnm` as freely variable inputs, and `lninv` as the proxy variable. We type

```
. set memory 96m

. use opreg

. xtset gvkey year
       panel variable:  gvkey (unbalanced)
        time variable:  year, 1995 to 2002, but with gaps
                delta:  1 unit

. gen firmid = gvkey

. sort firmid year

. by firmid: gen count = _N

. gen survivor = count == 8

. gen has95 = 1 if year == 2002

. sort firmid has95

. by firmid: replace has95 = 1 if has95[_n-1] == 1

. replace has95 = 0 if has95 == .

. sort firmid year

. by firmid: gen has_gaps = 1 if year[_n-1] != year-1 & _n != 1

. sort firmid has_gaps

. by firmid: replace has_gaps = 1 if has_gaps[_n-1] == 1

. replace has_gaps = 0 if has_gaps == .

. by firmid: generate exit = survivor == 0 & has95 == 0 & has_gaps != 1
> & _n == _N

. replace exit = 0 if exit == 1 & year == 2002

. opreg lny, exit(exit) state(age lnkop) proxy(lninv) free(lnl lnm) cvars(t)
> vce(bootstrap, seed(1) rep(250))
Bootstrap replications (250)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
  (output omitted)
..................................................   250
```

```
Olley-Pakes productivity estimator              Number of obs      =     19710
Group variable (i): gvkey                       Number of groups   =      3722
Time variable (t): year
                                                Obs per group: min =         1
                                                               avg =       5.3
                                                               max =         8

                                (Replications based on 3722 clusters in gvkey)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lny          |
         age |  -.0049433   .9959707    -0.00   0.996     -1.95701    1.947123
       lnkop |   .1608403   .0570089     2.82   0.005     .0491049    .2725758
         lnl |   .1573041   .0270405     5.82   0.000     .1043058    .2103024
         lnm |   .7243997   .0284511    25.46   0.000     .6686366    .7801628
           t |   -.016153   .0032608    -4.95   0.000    -.0225441    -.009762
------------------------------------------------------------------------------
State:          age lnkop
Free:           lnl lnm
Control:        t
Proxy:          lninv
```

Table 1 compares the OP results with OLS to examine whether controlling for these biases makes a difference in estimating the parameters of the production function. The estimated results follow our a priori expectations regarding the bias caused by simultaneity and selection problems. When these biases are not controlled for, the coefficients associated with variable inputs (e.g., labor and materials) are expected to have an upward bias, and the coefficients associated with quasi–fixed inputs (e.g., capital) are expected to be biased downward (Olley and Pakes 1996). As illustrated in table 1, the coefficients on variable inputs and quasi–fixed inputs move in a direction suggesting the elimination of these biases. Thus controlling for the biases from simultaneity and selection seems to be important because differences in the magnitudes of the coefficients arise.

Table 1. Production function estimates: OLS and OP estimation results

Variable	OP	OLS
Materials	0.724	0.739
	(0.028)***	(0.009)***
Labor	0.157	0.182
	(0.027)***	(0.010)***
Capital	0.161	0.141
	(0.057)***	(0.008)***
Age	−0.005	−0.006
	(0.996)	(0.001)***
Trend	−0.016	−0.017
	(0.003)***	(0.003)***

Standard errors in parentheses. Standard errors in OP model are bootstrapped using 250 replications. *** Significant at 1% level.

5 Conclusion

Researchers often estimate a production function to obtain a measure of firm productivity. Such estimates, however, may suffer from the presence of selection and simultaneity biases in the estimates of the input coefficients needed to construct a productivity measure. Olley and Pakes (1996) introduce a semiparametric estimator that controls for these biases when estimating production function parameters. This methodology allows us to obtain consistent estimates of the input coefficients and thus to obtain reliable productivity measures.

In this paper, we have provided a brief review of this methodology and have described the Stata command, `opreg`, that implements it. We have illustrated our code by using unbalanced panel data for firms in the COMPUSTAT North America database during the period 1995–2002. Our results show that, when simultaneity and selection biases are not controlled for, the coefficients associated with variable inputs are biased upward and the coefficient for the capital input is biased downward.

The findings indicate that, in order to obtain consistent estimates of the production function parameters and thus to obtain reliable productivity measures, one should not ignore the selection and simultaneity issues in the estimation of the production function.

6 Saved results

In addition to various results set by `bootstrap`, `opreg` saves in `e()`:

Scalars	
`e(Nprobit)`	number of observations used in first (`probit`) stage
`e(Nreg)`	number of observations used in second (`regress`) stage
`e(Nnl)`	number of observations used in third (`nl`) stage
Macros	
`e(cmdname)`	`opreg`
`e(title)`	Olley–Pakes regression
`e(dv1)`	stage 1 dependent variable
`e(dv2)`	stage 2 dependent variable
`e(free)`	variables specified in `free()`
`e(cvars)`	variables specified in `cvars()`
`e(proxy)`	variable specified in `proxy()`
`e(state)`	variables specified in `state()`
Matrices	
`e(b)`	coefficient vector
`e(V)`	variance–covariance matrix

7 Appendix

Because the Olley and Pakes (1996) procedure involves three steps and different implementations may differ in how each step is carried out, here we present a do-file that illustrates the version implemented by `opreg`. Running this do-file will yield the same point estimates that our command reports:

begin do-file

```
use opreg

xtset gvkey year

drop if missing(lninv)

// Create terms for polynomial in (i,k,a)
gen double lninvlnkop = lninv*lnkop
gen double lninvage = lninv*age
gen double lnkopage = lnkop*age
gen double lninvsq = lninv^2
gen double lnkopsq = lnkop^2
gen double agesq = age^2


// Step I - regress lny on variable inputs and
// polynomial in i, a, k
regress lny lnl lnm lninv lnkop age t lninvlnkop lninvage lnkopage  ///
        lninvsq lnkopsq agesq
predict double lny_hat if e(sample), xb
scalar b_lnl = _b[lnl]
scalar b_lnm = _b[lnm]


// Step II -- Estimate probability of survival
probit exit L.(lninv lnkop age t lninvlnkop lninvage lnkopage   ///
        lninvsq lnkopsq agesq)
predict phat if e(sample), pr

// Step III -- Nonlinear regression of y - lnl*b_lnl - lnm*b_lnm
// on age, capital, and the polynomial to control for selection

// First, get phi_hat
generate double phi_hat = lny_hat - lnl*b_lnl - lnm*b_lnm

// Next, generate the depvar for the nonlinear equation
// Output minus the contributions of the variable inputs
generate double lhs = lny - lnl*b_lnl - lnm*b_lnm

// mark out missing observations
generate useme = 1
gen l1phi = L.phi_hat
gen l1lnkop = L.lnkop
gen l1age = L.age

foreach var of varlist lhs lnkop age l1phi l1lnkop l1age {
        replace useme = 0 if `var´ >= .
}

gen double phat2 = phat^2

// Finally, fit the nonlinear model to get capital and age coefs.
nl ( lhs = b0 + bk*lnkop + ba*age +           ///
        t1*(l1phi - bk*l1lnkop - ba*l1age) +      ///
        t1sq*(l1phi - bk*l1lnkop - ba*l1age)^2 +      ///
        t2*phat + t2sq*phat^2 +             ///
        t1t2*(l1phi - bk*l1lnkop - ba*l1age)*phat )   ///
        if useme
```

end do-file

8 Acknowledgments

We would like to thank Steven Olley and an anonymous referee for their helpful suggestions. We are also grateful to Wayne Gray for sharing the deflators with us.

9 References

Arellano, M., and S. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277–297.

Arellano, M., and O. Bover. 1995. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68: 29–51.

Bartelsman, E., and W. Gray. 2001. NBER Productivity Database. Available at www.nber.org.

Blundell, R., and S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87: 115–143.

———. 2000. GMM estimation with persistent panel data: An application to production functions. *Econometric Reviews* 19: 321–340.

Griliches, Z., and J. Mareisse. 1998. Production functions: The search for identification. In *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Prisch Centennial Symposium*, 169–203. Cambridge: Cambridge University Press.

Levinsohn, J., and A. Petrin. 2003. Estimating production functions using inputs to control for unobservables. *Review of Economic Studies* 70: 317–342.

Marschak, J., and W. H. Andrews. 1944. Random simultaneous equations and the theory of production. *Econometrica* 12: 143–205.

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

Pavcnik, N. 2002. Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants. *Review of Economic Studies* 69: 245–276.

Petrin, A., B. P. Poi, and J. Levinsohn. 2004. Production function estimation in Stata using inputs to control for observables. *Stata Journal* 4: 113–123.

Wooldridge, J. 2005. On estimating firm-level production functions using proxy variables to control for unobservables. Mimeo: Michigan State University.

About the authors

Mahmut Yasar is an assistant professor of economics at the University of Texas Arlington. Rafal Raciborski is a graduate student in the department of political science at Emory University. Brian Poi is a senior economist at StataCorp.