European Journal of Operational Research 249 (2016) 784–788

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Short Communication

Notes on technical efficiency estimation with multiple inputs and outputs

Mike G. Tsionas*

Lancaster University Management School, LA1 4YX, UK

ARTICLE INFO

Article history:
Received 14 March 2015
Accepted 17 October 2015
Available online 30 October 2015

Keywords:
Efficiency
Least squares
Multiple-output production
Limited information maximum Likelihood

ABSTRACT

Collier, Johnson and Ruggiero (2011) deal with the problem of estimating technical efficiency using regression analysis that allows multiple inputs and outputs. This revives an old problem in the analysis of production. In this note we provide an alternative maximum likelihood estimator that addresses the concerns. A Monte Carlo experiment shows that the technique works well in practice. A test for homotheticity, a critical assumption in Collier, Johnson and Ruggiero (2011) is constructed and its behavior is examined using Monte Carlo simulation and an empirical application to European banking.

1. Introduction

To make things concrete, suppose we have multiple outputs in vector $\mathbf{y}_{it} \in \Re^M$, and there exists an aggregator function of the constant elasticity of substitution (CES) type:

$$g(\mathbf{y}_{it}) = (\alpha_1 y_{1,it}^{\rho} + \alpha_2 y_{2,it}^{\rho})^{1/\rho},\ i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{1}$$

although other aggregator can serve just as well. The model is:

$$g(\mathbf{y}_{it}) = f(\mathbf{x}_{it}) + v_{it} - u_{it}, \tag{2}$$

where $f(\mathbf{x}_{it})$ is a functional form that shows how inputs $\mathbf{x}_{it} = [x_{1,it}, \ldots, x_{K,it}]' \in \Re^K$ contribute to the production of aggregate output, v_{it} is a two sided error term and $u_{it} \geq 0$ represents technical inefficiency. For example, with a Cobb-Douglas functional form and K inputs, Eq. (2) is:

$$\ln g(\mathbf{y}_{it}) = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{k,it} + v_{it} - u_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T. \tag{3}$$

There are certain important issues in technical efficiency analysis with multiple outputs. Since outputs are jointly produced given the inputs, Eqs. (2) or (3) do *not* specify a joint system of equations for $\mathbf{y}_{it}$. In fact, with M outputs, $M-1$ equations are missing, as noted by Fernandez, Koop, and Steel (2002). Therefore, the aggregator function approach is not enough for a complete analysis of the problem.

It is a, relatively, common practice in the literature to assume, say,[1] $g(\mathbf{y}_{it}) = y_{1,it}^{\alpha_1} y_{2,it}^{\alpha_2}$ and use OLS in the following equation[2]:

$$\alpha_1 \ln y_{1,it} + \alpha_2 \ln y_{2,it} = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{k,it} + v_{it} - u_{it}, \tag{4}$$

or, alternatively, in obvious notation:

$$\ln y_{1,it} = \beta_0 + \gamma_0 \ln y_{2,it} + \sum_{k=1}^{K} \gamma_k \ln x_{k,it} + \tfrac{1}{\alpha_1}(v_{it} - u_{it}). \tag{5}$$

This equation shows plainly that we have an *endogenous* variable in the right-hand-side of the equation and, therefore, OLS *cannot* be used because of correlation with the error term, $v_{it} - u_{it}$. This is very often forgotten in applied research. This point applies not only to regressions involving aggregator functions but also to distance functions. Input- or output-oriented distance functions are homogeneous of degree one with respect to inputs and outputs, respectively. Therefore, imposition of homogeneity results in a form similar to (5).

[1] To the author's knowledge the only reason why a Cobb-Douglas aggregator may be undesirable is because it does not satisfy the second order conditions of profit maximization. A Cobb-Douglas aggregator is, however, consistent with cost minimization. Of course it is also undesirable in the sense that it is not flexible enough like the CES or translog functional forms. Here we use it for simplicity in presentation of the main points.

[2] This, and similar practices, are nicely reviewed in Kumbhakar and Lovell (2000, pp. 93–95) and the cited references.

* Tel.: +44 302108203338.
E-mail address: tsionas@otenet.gr, tsionas@aueb.gr

http://dx.doi.org/10.1016/j.ejor.2015.10.036
0377-2217/

The proper method of estimation is limited information maximum likelihood (LIML) *if* the inputs are exogenous and prices are not available[3]. However, this is also a problematic assumption.

Under specific behavioral assumptions, inputs are endogenous. For example under profit maximization or cost minimization, inputs are endogenously selected and outputs are, respectively, endogenous or predetermined. In effect, OLS is not the proper method of estimation. Using LIML requires predetermined variables like prices (which, by assumption, are not available.) However, lagged values of $\mathbf{x}_{it}$ and $\mathbf{y}_{it}$ can be used as they are predetermined from the point of view of period t, provided v_{it} and u_{it} are not autocorrelated. An alternative estimation technique is the generalized method of moments (GMM) with a one-sided error term and explicit correlation allowed between $\mathbf{x}_{it}$ and the error components (Tran & Tsionas, 2013).

Collier et al. (2011), CJR henceforth, deal with the problem of estimating technical efficiency using OLS regression analysis that allows multiple inputs and outputs. Specifically, technical efficiency can be estimated using regression models with multiple inputs and outputs *without* input price data. CJR propose to use DEA analysis in a first stage -without the input constraints since they assume separability from inputs- to derive an aggregate output measure, say S_{it}. The DEA problem solved in CJR is the following, for each observation "o":

$$\max:\ \Theta_o,$$
$$\sum_{j=1}^{NT} \Lambda_j y_{jm} \geq \Theta_o y_{om},\ m = 1, \ldots, M,$$
$$\sum_{j=1}^{NT} \Lambda_j = 1,$$
$$\Lambda_j \geq 0,\ j = 1, \ldots, NT. \tag{6}$$

Aggregate output is defined as:

$$S_o = \frac{1}{\Theta_o},\ o = 1, \ldots, NT. \tag{7}$$

Then, instead of (2) they propose regression analysis in the following model:

$$\log S_{it} = f(\mathbf{x}_{it}) + v_{it} - u_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T. \tag{8}$$

As $\mathbf{x}_{it}$ is often used in logs in (8) a Cobb-Douglas function has the following form:

$$\log S_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} - u_{it}\ i = 1, \ldots, N,\ t = 1, \ldots, T. \tag{9}$$

This methodology has been applied by Collier, Mamula, and Ruggiero (2014). Under homotheticity, this bypasses the problem of having an endogenous variable in the right-hand-side, and falls in between DEA and stochastic frontier analysis: It allows measurement error in the aggregate output, and statistical assumptions about the error terms v_{it} and u_{it}. However, it does not addresses concerns about the potential endogeneity of inputs, in which case $\mathbf{x}_{it}$ and v_{it} are correlated.

The main problem, however, is not to seek an output aggregator, since homotheticity cannot be taken for granted. In Fernandez et al. (2002) the problem in (3) is recognized for *what it truly is*, viz. several inputs $\mathbf{x}_{it}$ are used to *jointly produce* outputs $\mathbf{y}_{it}$. The problem is not which aggregator function must be used (a CES would be just fine for most purposes) or how to aggregate the outputs, but how to account for the *endogenous character* of $\mathbf{y}_{it}$ in this context. From the econometric point of view, the problem is that with M outputs there are M endogenous variables but only one equation, viz. (2) or (3). Therefore, there are $M-1$ missing equations to complete the system.

[3] If prices are available the system can be completed using the first order conditions from cost minimization or profit maximization. An alternative has been introduced by Atkinson and Tsionas (2015) where the first order conditions are used, and unobserved prices are treated as latent variables in the context of a Bayesian hierarchical model.

2. An alternative estimation technique

The question is how to deal with the problem when *both* outputs and inputs are endogenous and prices are *not* observed. To complete the model provided by (3) we consider the reduced form:

$$\begin{bmatrix} \tilde{\mathbf{y}}_{it} \\ \mathbf{x}_{it} \end{bmatrix} = \boldsymbol{\Pi}\mathbf{z}_{it} + \boldsymbol{\varepsilon}_{it}, \tag{10}$$

where $\tilde{\mathbf{y}}_{it} = [y_{2,it}, \ldots, y_{m,it}]'$, $\mathbf{z}_{it}$ is a vector of predetermined variables, $\boldsymbol{\varepsilon}_{it}$ is a vector of error terms and $\boldsymbol{\Pi}$ is a matrix of unknown parameters. Under the assumption that $u_{it} = u_i,\ \forall t = 1, \ldots, T$ are fixed parameters and:

$$[v_{it}, \boldsymbol{\varepsilon}_{it}']' \sim \mathcal{N}_M(\mathbf{0}, \boldsymbol{\Sigma}), \tag{11}$$

the system of (3) and (10) can be estimated using limited information maximum likelihood (LIML) along the lines suggested by Pagan (1979). The variables in $\mathbf{z}_{it}$ include *only* firm and time dummies so that it is not necessary to think about the possibility of other instruments, which may not be available at all in practice. For example, the use of lagged values of $\tilde{\mathbf{y}}_{it}$ and $\mathbf{x}_{it}$ is often problematic if they are only weakly correlated with the endogenous variables.

As in CJR, we assume time-invariant technical inefficiency, that is $u_{it} = u_i, \forall t = 1, \ldots, T$. As the inclusion of firm-specific effects in (10) prevents the identification of u_is we use the *nonlinear transformation*:

$$u_i = \exp\left(-\varphi_i^2\right),\ \forall i = 1, \ldots, n, \tag{12}$$

where the φ_is are unrestricted. With this transformation, the u_is are *always positive and less than one*, thus making it unnecessary to apply a corrected ordinary least squares (COLS) transformation, viz. $\hat{u}_i = u_i - \max_{i=1,\ldots,n} u_i$, to obtain technical inefficiencies. To the author's knowledge, this transformation has not been used before although it has considerable merit, at the cost of requiring nonlinear estimation techniques.

3. Monte Carlo evidence

To see specifically how the problems mentioned in the previous section can be addressed, we consider a model using (1) with $\rho = \frac{1}{2}$. The first output is are generated from a standard lognormal distribution. Next, we generate $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$. The three inputs (in log terms) are generated as follows:

$$\begin{aligned} x_{1,it} &= \alpha v_{it} + \mathbf{D}_{it}\boldsymbol{\gamma}_1 + \xi_{1,it}, \\ x_{2,it} &= x_{1,it} + \mathbf{D}_{it}\boldsymbol{\gamma}_2 + \xi_{2,it}, \\ x_{3,it} &= x_{1,it} + \mathbf{D}_{it}\boldsymbol{\gamma}_3 + \xi_{3,it}, \end{aligned} \tag{13}$$

where $\xi_{j,it} \sim N(0, 1)$. Therefore, inputs are mutually correlated as well as correlated with v_{it}. The correlation coefficient between $x_{1,it}$ and v_{it} is $\varrho = \frac{\alpha\sigma_v^2 + 1}{\sigma_v\sqrt{\alpha^2\sigma_v^2 + 1}}$. Given σ_v we can vary α so that we obtain different values of ϱ. Moreover, $\mathbf{D}_{it}$ is a vector of firm and time dummies and $\boldsymbol{\gamma}_j$ are respective coefficients, which are generated from a uniform distribution in [1, 1]. We have $N+T-1$ coefficients in each input equation.

The model is the same as in (3) with constant term -1 and slope coefficients equal to $\frac{1}{3}$. Finally, the second output is generated from (2) where $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$ and independently $u_{it} \sim \mathcal{N}_+(0, \sigma_u^2)$. Denote $\sigma = \sqrt{\sigma_v^2 + \sigma_u^2}$ and $\lambda = \frac{\sigma_u}{\sigma_v}$. For practical reasons, we can set $\sigma = 0.3$ and $\lambda = 1$ which is a typical case in empirical studies and examine the rank correlation between true and estimated inefficiencies for various values of ϱ. The number of firms is N, the number of time periods is T and we assume $u_{it} = u_i,\ \forall t = 1, \ldots, T$.

LIML is implemented using a standard conjugate-gradients algorithm without analytic derivatives and we consider 10,000 alternative data sets. Our evidence is summarized in Table 1. Clearly, the