Nonparametric Least Squares Methods for Stochastic Frontier Models

Léopold Simar, Ingrid Van Keilegom, Valentin Zelenyuk*

* School of Economics and Centre for Efficiency and Productivity Analysis (CEPA), The University of Queensland, Australia

APPC 2014

The Background

In productivity and efficiency analysis, researchers are primarily interested in **two aspects**:

- the estimation of a function characterizing the **production frontier and its characteristics** (marginal productivity, elasticities, etc.),
- **explaining of variation in inefficiency**.

One of the most popular approaches for studying these aspects is referred to as **stochastic frontier analysis (SFA)**, introduced by Aigner, Lovell and Schmidt (1977) and Meusen and van den Broek (1977) (henceforth ALSMB).

The SFA paradigm has a **very appealing feature** relative to other methods—it allows the presence of both an **inefficiency** term modeling the distance of an observation to the frontier and the more traditional **error** term (as in most regression models) allowing for noise.

The Background (cont.)

Here we work with **SFA** paradigm but in a **non-parametric and semi-parametric** context. Specifically, the model is

$$y = m(x, z) - u + v, \tag{1}$$

where $y \in \mathbb{R}_+$ represents the output that can be produced with the inputs $x \in \mathbb{R}_+^p$ in the presence of heterogeneous conditions $z \in \mathbb{R}_+^q$, using the available technology characterized by the production frontier $m(\cdot, \cdot)$; the output y is adjusted by some possible inefficiency level $u \in \mathbb{R}_+$ and by some statistical noise $v \in \mathbb{R}_+$

The two terms u and v are unobserved random variables which may vary with the inputs x as well as with the other variables z.

Estimation of Average Production Function

A relatively easy task is to estimate the '**average production function**' relationship (i.e., if inefficiency term u is ignored), and this can be done in a **fully non-parametric** way.

Letting $\varepsilon = v - u + E(u|x,z)$, and $r_1(x,z) = m(x,z) - E(u|x,z)$ we can rewrite (1) as

$$y = r_1(x,z) + \varepsilon \tag{2}$$

Since $E(\varepsilon|x,z) = 0$, and $V(\varepsilon|x,z) = V_v(x,z) + V_u(x,z) \in (0,\infty)$, we can apply any valid non-parametric estimator to estimate $r_1(x,z) = E(y|x,z)$, e.g., **local polynomial least squares (LPLS)** estimator, which is a fully non-parametric approach with good asymptotic properties and is relatively easy and fast to compute.

In one estimation LPLS can produce consistent and **asymptotically normal estimates** of $r_1(x,z)$, which we will denote here as $\widehat{r}_1(x,z)$, and estimates of its k^{th}-order partial derivatives (if the order of the polynomial is chosen to be at least of order k).

LPLS Est. of Average Production Function

From i.i.d. data $\{(Y_i, X_i, Z_i) \ : \ i = 1, ...n\}$ we can estimate the local linear estimate of $r_1(x, z)$ by solving, for any given value (x, z)

$$(\alpha_{x,z}, \beta_{x,z}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left[Y_i - (\alpha + \beta'(W_i - w))\right]^2 K_h\big((W_i - w)/h\big), \tag{3}$$

where $W_i = (X_i, Z_i)$, $w = (x, z)$ and h denotes the bandwidths (with some abuse of notations). Then we have

$$\begin{aligned} \widehat{r}_1(x, z) &= \alpha_{x,z} \\ \widehat{\nabla} r_1(x, z) &= \beta_{x,z}, \end{aligned}$$

where the second equation provides an estimate of the gradient of $r_1(x, z)$ at (x, z).

Est. of Average Production Function

Note that $\widehat{r}_1(x,z)$ will be **an estimate not of the production frontier** $m(x,z)$ but of $r_1(x,z) = m(x,z) - E(u|x,z)$, which we refer to as the average production function.

Since $E(u|x,z) \geq 0$, we have $r_1(x,z) \leq m(x,z)$, $\forall x \in \mathbb{R}^p_+$, $\forall z \in \mathbb{R}^d_+$. Thus, clearly, $r_1(x,z) = m(x,z)$, $\forall x \in \mathbb{R}^p_+$, $\forall z \in \mathbb{R}^d_+$ if and only if $E(u|x,z) = 0$, i.e., if and only if there is no inefficiency.

Hence, if there is some inefficiency, then $\widehat{r}_1(x,z)$ would be a downward-biased estimator of $m(x,z)$ at any level of inputs.

Moreover, the **bias** is $E(u|x,z)$, and so it **is varying** with (x,z), unless $E(u|x,z) = E(u)$.

Note, however, that **some important information about the average production relationship** (e.g., the marginal productivity of inputs, the scale elasticity, etc.) is still contained in $r_1(x,z)$ and so can be inferred from $\widehat{r}_1(x,z)$.

Three Moments of the Total Error

Without specifying a particular choice for the local distributions of u and of v, we can also estimate the moments of ε since under the symmetry assumption on v,

$$E(\varepsilon|x,z) = 0, \tag{4}$$

$$E(\varepsilon^2|x,z) = V_v(x,z) + V_u(x,z) > 0, \tag{5}$$

$$E(\varepsilon^3|x,z) = -E\left[\left(u - E(u|x,z)\right)^3|x,z\right] \leq 0. \tag{6}$$

We will extend the idea of the Modified OLS (MOLS), originated in the full parametric, homoskedastic stochastic frontier models (see Olson et al., 1980) to our semi-parametric setup.

The idea is to exploit the fact that the residuals in (2) may help estimating the conditional moments of ε.

Estimation of 2nd and 3rd Moments

So, the residuals can be evaluated at any data point, and in particular, to obtain

$$\widehat{\varepsilon}_i = Y_i - \widehat{r}_1(X_i, Z_i),\ i = 1, ..., n. \tag{7}$$

It is known that for all i, conditionally on (Y_i, X_i, Z_i), $\widehat{\varepsilon}_i \to \varepsilon_i$. The same is true for the powers of ε.

Therefore, **consider the following regressions**:

$$\widehat{\varepsilon}^2 = E(\widehat{\varepsilon}^2|x, z) + e_2 = r_2(x, z) + e_2, \tag{8}$$

$$\widehat{\varepsilon}^3 = E(\widehat{\varepsilon}^3|x, z) + e_3 = r_3(x, z) + e_3, \tag{9}$$

where by construction $E(e_2|x, z) = 0$ and $E(e_3|x, z) = 0$.

Using a nonparametric technique (e.g., LPLS), the regression functions $r_2(x, z)$ and $r_3(x, z)$ can be **consistently estimated** from $\{(\widehat{\varepsilon}_i^2, X_i, Z_i)|i = 1, ..., n\}$ and $\{(\widehat{\varepsilon}_i^3, X_i, Z_i)|i = 1, ..., n\}$.

Estimation of 2nd and 3rd Moments

We prove that $\widehat{r}_2(x,z)$ and $\widehat{r}_3(x,z)$ are **consistent and asymptotically normal estimates** of $E(\varepsilon^2|x,z)$ and $E(\varepsilon^3|x,z)$ respectively.

Now from (5)–(6), note the link with conditional moments of u and v, so plugging these estimates in the equations will provide information on the moments of u and v.

In order to identify the frontier level and some important parameters of the model, e.g., $m(x,z)$, $E(u|x,z)$, $V_u(x,z)$ and $V_v(x,z)$ we **need to select local parametric assumptions** for both the density of $u|x,z$ and of $v|x,z$.

From our discussion above it must be clear that for implementing this task we need to have information on $E(u|x,z)$ and this is where (for the cross-sectional data framework) we **need to make local parametric assumptions on the types of distributions** of $u|x,z$ and $v|x,z$.

$$(v|x,z) \sim N(0,\sigma_v^2(x,z)),\ v \in (-\infty,\infty), \tag{10}$$

$$(u|x,z) \sim |N(0,\sigma_u^2(x,z))|,\ u \in (0,\infty), \tag{11}$$

where we also assume that, conditionally on (x,z), u and v are independent.

As a result, we would have

$$E(u|x,z) = \sqrt{\frac{2}{\pi}}\sigma_u(x,z) \tag{12}$$

$$E(\varepsilon^2|x,z) = V(\varepsilon|x,z) = \sigma_v^2(x,z) + \left(\frac{\pi-2}{\pi}\right)\sigma_u^2(x,z) \tag{13}$$

$$E(\varepsilon^3|x,z) = \sqrt{\frac{2}{\pi}}\left(1-\frac{4}{\pi}\right)\sigma_u^3(x,z). \tag{14}$$

Rearranging the system of equations given in (13)-(14), and solving it for $\sigma_u^3(x,z)$ and $\sigma_v^2(x,z)$ we get

$$\sigma_u^3(x,z) = \sqrt{\frac{\pi}{2}}\left(\frac{\pi}{\pi-4}\right)E(\varepsilon^3|x,z) \tag{15}$$

$$\sigma_v^2(x,z) = E(\varepsilon^2|x,z) - \left(\frac{\pi-2}{\pi}\right)\sigma_u^2(x,z) \tag{16}$$

The LPLS estimates $\widehat{r}_2(x,z)$ and $\widehat{r}_3(x,z)$ are asymptotically equivalent to the corresponding true conditional moments of ε, so we can use them to **get consistent estimates of the conditional variances** at each point of interest (x,z), i.e.,

$$\widehat{\sigma}_u^3(x,z) = \sqrt{\frac{\pi}{2}}\left(\frac{\pi}{\pi-4}\right)\widehat{r}_3(x,z) \tag{17}$$

$$\widehat{\sigma}_v^2(x,z) = \widehat{r}_2(x,z) - \left(\widehat{\sigma}_u^3(x,z)\right)^{2/3}\left(\frac{\pi-2}{\pi}\right) \tag{18}$$

Using these estimates, we can obtain the estimates of efficiency scores for each observation, e.g., by using the method of **Jondrow et al (1982)**—after **generalizing it to the heteroskedastic case**, by estimating $E(u_i|\varepsilon_i, x_i, z_i)$ instead of $E(u_i|\varepsilon_i)$.

Estimation of Average Production Function

Also, a **useful information** can be inferred from a consistent estimate of the **conditional mean of inefficiency** term, conditional on (x, z),

$$\widehat{E}(u|x,z) = \sqrt{\frac{\pi}{2}}\left(\sqrt{\frac{\pi}{2}}\frac{\pi}{\pi-4}\widehat{r}_3(x,z)\right)^{1/3} \tag{19}$$

Furthermore, estimates of $\widehat{E}(u|x,z)$ at every combination of interest (x,z) can then be used to recover a consistent estimate of the stochastic frontier, $m(x,z)$, via

$$\widehat{m}(x,z) := \widehat{r}_1(x,z) + \widehat{E}(u|x,z). \tag{20}$$

Asymptotic properties of $\widehat{m}(x,z)$, $\widehat{E}(u|x,z)$, $\widehat{\sigma}_u(x,z)$ and $\widehat{\sigma}_v(x,z)$ are inherited from the asymptotic properties of $\widehat{r}_1(x,z)$, $\widehat{r}_2(x,z)$ and $\widehat{r}_3(x,z)$ and we provide details in appendix.

Robust Analysis of Determinants of Inefficiency

In the traditional SFA setup, when **statistical noise is symmetric** while **inefficiency term is asymmetric**, all the information about the inefficiency is essentially contained in the negative skewness of the composite error.

So, studying inefficiency boils down into **studying the conditional skewness** with respect to (x, z), which can be done via a **non-parametric** regression approach.

Interestingly, to perform such analysis, we do not need to specify the distribution of $v|x, z$: we can use any **one-parameter scale family for the density** $u|x, z$, without specifying which member of the family is chosen.

The density $f_u(u|x,z)$ belongs to the **one parameter scale family** if

$$f_u(u|x,z) = \frac{1}{\sigma_u(x,z)} g\left(\frac{u}{\sigma_u(x,z)}\right), \tag{21}$$

where $g(\cdot)$ is any density on $\mathbb{R}_+$. Examples of this are the exponential, the half-normal, the gamma with fixed shape parameter, etc. In this family it is easy to show that for all $j \geq 1$

$$E(u^j|x,z) = \sigma_u^j(x,z) k_j, \tag{22}$$

as long as the j^{th} moment of g, $k_j = \int_0^\infty v^j g(v)\, dv$, exists. As a result, we also have

$$E(\varepsilon^3|x,z) = E\left[\left(u - E(u|x,z)\right)^3 \middle|\, x,z\right] = c\sigma_u^3(x,z). \tag{23}$$

where $c = (k_3 - 3k_2k_1 + 2k_1)$ is a constant (depends on g).

Robust Analysis of Determinants of Inefficiency

Often, practitioners are actually more interested in the **determinants of the inefficiency** rather than the inefficiency or the frontier per se.

Sometimes, researchers are even satisfied with at least the direction (sign) of the influence, although perhaps ideal information would be about the **elasticities of the inefficiency** w.r.t. certain variables, since they do not depend on units of measurement of the variables involved.

Let ψ_l be an element of (x, z), then the (partial) **elasticity measure** of $E(u|x, z)$ w.r.t. ψ_l, denoted by $\xi_{u/\psi_l}(x, z)$, is

$$\xi_{u/\psi_l}(x, z) := \frac{\partial E(u|x, z)}{\partial \psi_l} \frac{\psi_l}{E(u|x, z)} \tag{24}$$

assuming that $E(u|x, z) \neq 0$. Using (22) with $j = 1$ we immediately get

$$\xi_{u/\psi_l}(x, z) = \frac{\partial \sigma_u(x, z)}{\partial \psi_l} \frac{\psi_l}{\sigma_u(x, z)} \tag{25}$$

if $\sigma_u(x, z) \neq 0$ (i.e., if there is some inefficiency at (x, z)).

Although $\partial \sigma_u(x, z)/\partial \psi_l$ is **not directly estimable, we can still recover it** from estimate of $E(\varepsilon^3|x, z)$.

Indeed, using (22) when $j = 3$ we get

$$\begin{aligned} \frac{\partial E(\varepsilon^3|x,z)}{\partial \psi_l}\frac{\psi_l}{E(\varepsilon^3|x,z)} &= 3c\sigma_u^2(x,z)\frac{\partial \sigma_u(x,z)}{\partial \psi_l}\frac{\psi_l}{c\sigma_u^3(x,z)} \quad (26) \\ &= 3\xi_{u/\psi_l}(x,z) \end{aligned}$$

Therefore, we can **express the elasticity of inefficiency only in terms of the third moment of the total error**, i.e.,

$$\xi_{u/\psi_l}(x,z) = \frac{1}{3}\frac{\partial E(\varepsilon^3|x,z)}{\partial \psi_l}\frac{\psi_l}{E(\varepsilon^3|x,z)} \quad (27)$$

So, a non-parametric estimate of $\xi_{u/\psi_j}(x,z)$ can be obtained by replacing the true moment $E(\varepsilon^3|x,z)$ and $\partial E(\varepsilon^3|x,z)/\partial\psi_l$ with their **non-parametric estimates**, i.e., as

$$\hat{\xi}_{u/\psi_j}(x,z) = \frac{1}{3}\frac{\widehat{\partial r_3}(x,z)}{\partial\psi_l}\frac{\psi_l}{\hat{r}_3(x,z)} \tag{28}$$

where $\hat{r}_3(x,z)$ and $\widehat{\partial r_3}(x,z)/\partial\psi_l$, $l=1,...,p+q$ are, for example, the **LPLS estimates** of (9), provided, of course, that $\hat{r}_3(x,z)\neq 0$ for the particular combination of interest (x,z).

See appendix for a proof of consistency and asymptotic normality of the estimator $\hat{\xi}_{u/\psi_l}(x,z)$

Testing of Existance of Inefficiency

We would expect $\widehat{r}_3(x,z)$ being negative and significantly different from zero at some ranges of (x,z) where firms in the sample have significant inefficiency and insignificantly different from zero for some ranges of (x,z) where there is no significant inefficiency.

So, **existing tests developed for LPLS framework can be adapted** to test whether $\widehat{r}_3(x,z)$ is significantly different from zero or not.

In particular, it is well known that the LPLS estimator of a regression function (and of its derivatives) is asymptotically normally distributed (under some regularity conditions).

So, for a combination of interest (x,z), the null hypothesis about no inefficiency at this particular (x,z), i.e., $H_0 : r_3(x,z) = 0$, would be rejected in favor of the alternative hypothesis that inefficiency at (x,z) is present, i.e., $H_0 : r_3(x,z) < 0$, if the test statistic is beyond the critical value corresponding to the chosen significance level.

Testing of Determinants of Inefficiency

LPLS also allows inference about the **sign and size of the impact** of x and z on the inefficiency, by using estimates of $\nabla_x r_3(x,z)$ and $\nabla_z r_3(x,z)$, respectively, and testing their significance from zero, at a particular combination (x,z).

Specifically, the null hypothesis about no impact on inefficiency by a potential factor ψ_i, i.e., $H_0 : \xi_{u/\psi_j}(x,z) = 0$, where ψ_i is an element of (x,z), would be rejected in favor of the alternative hypothesis $H_1 : \xi_{u/\psi_j}(x,z) \neq 0$ if the statistic is beyond the critical value corresponding to the chosen significance level.

In practice, bootstrap-based inference adapted to LPLS, about $r_3(x,z) = 0$ or $\xi_{u/\psi_j}(x,z) = 0$, might give more accurate results than the inference based on asymptotic normality results for our estimator.

Emperical Example

See our Working Paper (on CEPA and ISBA websites) for details