Stochastic Frontier Analysis

Tristan D. Skolrud

EconS 504/EconS 513

March 21, 2016
Part 1

Introduction

Prior to ALS (Aigner, Lovell, and Schmidt 1977) and MvdB (Meeusen and Van den Broeck 1977), the estimation of parametric production functions started with the theoretical representation of a production function

\[ y_i = f(x_i, \beta) \]

where \( y_i \) represents the *maximal* amount of output \( y_i \) obtainable from inputs \( x_i \) and production technology \( f(x_i, \beta) \).
Estimation followed mathematical programming techniques (Aigner and Chu 1968), maximizing

\[ \sum_{i=1}^{n} (y_i - f(x_i, \beta)) \]

or

\[ \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \]

s.t. \( y_i \leq f(x_i, \beta) \).

Raises two questions:

1. How does one explain differences in \( y_i \) for identical \( x_i \)?
2. What accounts for a firm producing below (or above) the \( f(x_i, \beta) \) frontier?
Answer: *Measurement error*

But this fails to address the stochastic nature of production, long realized by economists and highlighted by the pioneering theoretical work of Farrell (1957).

ALS and MvdB sought to operationalize the theoretical framework of Farrell (1957), allowing for the estimation of a stochastic production frontier, where firms could operate below the frontier for two reasons:

1. Technical inefficiency
2. Statistical noise (measurement error)

How is technical inefficiency defined?
Technical Inefficiency
Intuitively, technical inefficiency is the amount by which all inputs can be proportionally reduced without a reduction in output.

[Graph]
Stochastic Frontier
With the idea of technical inefficiency in mind, consider the following parametric equation:

\[ y_i = f(x_i, \beta) + \varepsilon_i \]

where \( \varepsilon_i = v_i - u_i \) for \( i = 1, \ldots, n \) (firms) and

- \( v_i \) is a symmetric error term accounting for statistical noise
- \( u_i \) is a non-negative term accounting for technical inefficiency

Each firm’s output must lie on or below its frontier, \( y_i \leq f(x_i, \beta) + v_i \), which can vary randomly across firms or over time.
Checking for the initial presence of TE

- Observe that if \( u_i = 0 \), then \( \varepsilon_i = v_i \), implying that the error term is symmetric, which does not support the presence of technical inefficiency.
- However, if \( u_i > 0 \), then \( \varepsilon_i \) should be negatively skewed.
- SFA should start with a simple test (Schmidt and Lin 1984) of the presence of TE in the data. Consider the test statistic:

\[
(b_1)^{(1/2)} = \frac{m_3}{(m_2)^{(3/2)}}
\]

where \( m_2 \) and \( m_3 \) are the second and third sample moments of the OLS residuals of the previous model. \( m_3 < 0 \) indicates technical efficiency may be present, and \( m_3 > 0 \) is a sign that your model may be misspecified.
Checking for the initial presence of TE
As a quick note, the distribution for \((b_1)^{1/2}\) is not widely distributed, so it’s more common to test the statistic:

\[
b_{alt} = \frac{m_3}{(6m_2^3/n)^{1/2}} \sim N(0, 1)
\]
Estimation
Maximum likelihood is the preferred technique, representing an increase in efficiency over OLS. Of course, that means we require a variety of assumptions about the standard errors:

\[
E(v_i) = 0 \\
E(v_i^2) = \sigma_v^2 \\
E(v_i v_j) = 0 \text{ for all } i \neq j \\
E(u_i^2) = \text{constant} \\
E(u_i u_j) = 0 \text{ for all } i \neq j
\]

(“Corrected” OLS (COLS), GMM, and Bayesian methods have been used as well)
For maximum likelihood, we require parametric assumptions about the two disturbance terms. ALS use a normal distribution for the symmetric disturbance and a half-normal distribution for the technical inefficiency term:

\[ v_i \sim iid \ N(0, \sigma_v^2) \]
\[ u_i \sim iid \ N^+(0, \sigma_u^2) \]

Other popular choices for the inefficiency term are:

1. Truncated normal
2. Exponential
3. Gamma

In practice, half-normal is the default choice, and the remaining distributions are often used as robustness checks.
Half-normal density

- Negative values set to zero, positive values follow the right-half of a normal distribution
Half-normal density

- Note that the parameters $\mu$ and $\sigma^2$ in the half-normal distribution $N^+(\mu, \sigma^2)$ are *not* the mean and variance!
- The density is given by

$$f(x; \sigma) = \frac{\sqrt{2}}{\sigma \sqrt{\pi}} \exp \left( -\frac{x^2}{2\sigma^2} \right)$$

- The mean is

$$E(x) = \frac{\sigma \sqrt{2}}{\sqrt{\pi}}$$

- The variance is

$$V(x) = \sigma^2 \left( 1 - \frac{2}{\pi} \right)$$

- And the density has support over all $x \in [0, \infty)$
Reparameterization

Reparameterize variance terms by defining $\gamma = \sigma_u^2/\sigma^2$, where $\sigma^2 = \sigma_u^2 + \sigma_v^2$. Benefits:

- Reduces search area of $\gamma$, $\{\gamma \in (0, 1)\}$
- Easy interpretation: $\gamma \to 1$ implies more of the variation is attributed to inefficiency, and $\gamma \to 0$ implies more of the variation due to statistical noise
Likelihood

With the reparameterization, Battese and Corra (1977) demonstrate that the log-likelihood function can be written:

$$\ln L = -\frac{n}{2} \ln \left( \frac{\pi}{2} \right) - \frac{n}{2} \ln(\sigma^2) + \sum_{i=1}^{n} \ln(1 - \Phi(z_i)) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i \beta)^2$$

where

$$z_i = \frac{y_i - x_i \beta}{\sigma} \sqrt{\frac{\gamma}{1 - \gamma}}$$

Recall the rule that the density of a sum of random variables, $f(z)$, where $Z = X + Y$ and $f(x)$ and $g(y)$ are the resp. densities, is given by the convolution

$$(f \ast g)(z) = \int_{-\infty}^{\infty} f(z - y)g(y)dy$$
Algorithm

Estimation of the stochastic frontier follows a three-step algorithm:

1. Obtain OLS estimates from \( y_i = f(x_i, \beta) + \nu_i \)
2. Adjust intercept \( \beta_0 \) and \( \sigma^2 \) for bias, and iterate \( \gamma \in (0, 1) \) over the likelihood function to identify a preferred starting value.

\[
\hat{\sigma}^2 = \frac{n - k}{n} \left( \frac{\pi}{\pi - 2\gamma} \right)
\]

\[
\hat{\beta}_0 = \hat{\beta}_0(OLS) + \sqrt{\frac{2\gamma\hat{\sigma}^2}{\pi}}
\]

3. Use the values from step 2 as the starting values in a \( k + 2 \) dimensional nonlinear maximization problem.
Firm-Level Technical Efficiency Estimates
Most common output oriented measure of technical efficiency is the ratio of observed output to the corresponding stochastic frontier output (Coelli et al. 2005):

\[ TE_i = \frac{q_i}{f(x_i, \beta) + v_i} = \frac{f(x_i, \beta) + v_i - u_i}{f(x_i, \beta) + v_i} \]

When the dependent variable is logged (CD, TL)*, \( TE_i \) reduces to the convenient:

\[ TE_i = \exp(-u_i) \]

*I am unaware of any study that does not utilize a logged dependent variable in SFA*
Estimator for $TE_i$

There are several estimators of $TE_i$ based on the previous derivation (c.f. Jondrow et al. 1982). One of the more popular forms was developed by Battese and Coelli (1988), who used the conditional density $p(u_i|q_i)$ to derive

$$\hat{TE}_i = E(\exp(-u_i)|q_i) = \left[ \Phi \left( \frac{u_i^*}{\sigma_*} - \sigma_* \right) / \left( \frac{u_i^*}{\sigma_*} \right) \right] \exp \left( \frac{\sigma_*^2}{2} - u_i^* \right)$$

where $u_i^* = -(\ln q_i - x_i^\prime \beta)\hat{\sigma}_u^2/\hat{\sigma}^2$, and $\hat{\sigma}_*^2 = \hat{\sigma}_v^2 \hat{\sigma}_u^2 / \hat{\sigma}^2$. Note that

$$\hat{\sigma}_u^2 / \hat{\sigma}^2 = \hat{\gamma}$$

$$\hat{\sigma}_*^2 = \hat{\sigma}^2 \hat{\gamma}(1 - \hat{\gamma})$$
Part 2

Introduction

- The true nature of production is stochastic, especially in agriculture.
- The authors suspect that increased instances of drought, higher average temperatures, and hotter daily maximums may be decreasing technical efficiency in livestock operations, particularly dairies.
- The authors specify a model wherein technical efficiency \textit{and} a vector of variables suspected to influence technical efficiency (associated with climate) are estimated simultaneously.
- Results indicate that a one unit increase in the annual THI (temperature-humidity index) load is associated with a 3.7 percent reduction in output.
- The question for us is: how did they figure this out?
Estimation Strategy

- **Objective**: Estimate the impact of THI load on technical efficiency
- **Starting point**: ALS (1977)/MvdB (1977)

\[
\ln(q_i) = f(x_i, \beta) + v_i - u_i
\]

(where \( f(x_i, \beta) \) is parameterized as Translog)

- Recall the deterministic frontier is \( f(x_i, \beta) \), the stochastic frontier is \( f(x_i, \beta) + v_i \), where \( v_i \) is a symmetric random shock, and \( u_i \geq 0 \) represents inefficiency

- With a logged dependent variable, technical efficiency is represented by

\[
TE_i = \frac{q_i}{\exp(f(x_i, \beta) + v_i)} = \exp(-u_i)
\]

which varies between 0 and 1, where \( TE_i = 1 \) indicates perfect technical efficiency
Estimation

- Assume default normal/half-normal error specification, define \( y_i = \ln(q_i) \) and \( f(x_i, \beta) = x_i \beta \), parameterize the log-likelihood function as

\[
\ln L(y_i|\beta, \sigma, \lambda) = \sum_{i=1}^{n} \left( \frac{1}{2} \ln \left( \frac{2}{\pi} \right) - \ln \sigma + \ln \Phi(-w_i) - \frac{\varepsilon_i^2}{2\sigma^2} \right)
\]

where

\[
\sigma^2 = \sigma_u^2 + \sigma_v^2 \\
\lambda = \sigma_u / \sigma_v \\
\varepsilon_i = y_i - x_i \beta \\
w_i = \varepsilon_i \lambda / \sigma
\]

and \( \Phi(\bullet) \) is the standard normal cumulative distribution function.
Estimation

Key and Sneeringer employ the Jondrow et al. (1982) version of the expectation of $u_i$ conditional on $\varepsilon_i$:

$$E(u_i|\varepsilon_i) = \frac{\sigma \lambda}{1 + \lambda^2} \left( \frac{\phi(w_i)}{1 - \Phi(w_i)} - w_i \right)$$

- With this estimate of $u_i$, how does one calculate the impact of a set of exogenous factors on its determination?
- Two-step estimation? Just estimate the $u_i$’s as normal, and then use it as a dependent variable in a second-stage estimation, regressed on factors thought to have influence
- **No.** Results in biased and inefficient estimates (Wang and Schmidt 2002)
Estimation
A more robust alternative to estimate technical efficiency along with the factors that influence it in a single step. To do this:

- Define the variance of the underlying half-normal distribution of $u_i$, $\sigma^2_u$, as a function of observable factors $z_u$ and a set of parameters $\delta_u$:

$$\sigma^2_{ui} = \exp(z_{ui}\delta_u)$$

- With this formulation, the factors in $z_{ui}$ directly impact the mean and variance of the inefficiency term $u_i$, and subsequently, the estimate of technical efficiency (still use Jondrow et al. 1982)
Estimation
Note: This formulation increases the dimensionality of the nonlinear maximization problem by the size of the $\delta_u$ vector. The likelihood function is now

$$\ln L(y_i|\beta, \sigma, \lambda, \delta_u) = \sum_{i=1}^{n} \left( \frac{1}{2} \ln \left( \frac{2}{\pi} \right) - \ln \sigma + \ln \Phi(-w_i) - \frac{\varepsilon_i^2}{2\sigma^2} \right)$$

where

$$\sigma^2 = \exp(z_{ui} \delta_u) + \sigma_v^2$$
$$\lambda = \exp(z_{ui} \delta_u) \sigma_v$$
$$\varepsilon_i = y_i - x_i \beta$$
$$w_i = \varepsilon_i \lambda / \sigma$$
What did Key and Sneeringer find?

- Postulated the impact of THI load, operator education, operator age, operator experience, operation size, and a measure of specialization

- THI = (dry bulb temperature in degrees celsius) + (0.36 × dew point temperature) + 41.2.

- THI load is a measure of the duration and extent above this threshold

- Results: THI load has a large, significant impact on technical efficiency in dairy production. Using 2010 estimates, inefficiency loss from heat stress reduces value by approximately $1.2 billion/year.

- Climate change simulations: Lost production could get much, much worse depending on the climate simulation model used.