The Econometric Approach to Efficiency Analysis

William Greene
Stern School of Business
New York University
The Measurement of Productive Efficiency

Techniques

The Red Book
2nd Edition

Harold O. Fried
C. A. Knox Lovell
Shelton S. Schmidt, Editors

1993

2008
Stochastic Frontier Analysis

Subal C. Kumbhakar
and C. A. Knox Lovell
Essential Theory

♦ The Stochastic Frontier Model
♦ Panel Data Models: Fixed and Random Effects and Time Varying Inefficiency
♦ Linking Demand Systems and Cost Functions
♦ Decomposing Cost Inefficiency
♦ Profit Efficiency
♦ Shadow Prices
♦ Exogenous Influencies on Inefficiency
♦ Productivity and Technical Change
Surveys of Econometric Methods in Efficiency Analysis

Journal of Econometrics Annals Issues:
46 (1990): Frontier Analysis: Parametric and Nonparametric Approaches
121 (2004): Georgia Workshop

Greene, 1997, Frontier Production Functions

Bauer, 1990, Recent Developments in Econometric…

Murillo-Zamorano, 2004, Economic Efficiency and Frontier Techniques
The Literature is Large

- Special issues of *JPA* (Conference volumes)
- *JPA*, regular methods and pedagogy
- Other journals: *Journal of Applied Econometrics, Empirical Economics*, etc.
- 6 entries on *Journal of Econometrics* All Star list of 50 papers since 1980.
Inefficiency

Technical and Allocative Inefficiency

Koopmans (1951), Debreu (1951), Shephard (1953), Farrell (1957), ...
Econometrics and Inefficiency

\[ f(x=\text{inputs}, \beta, v=\text{noise}) \]

\[ u=\text{inefficiency} \]

Recurring Econometric Themes in Recent Literature

- The ALS Stochastic Frontier Model
  - Parametric formulations
  - Non- and semiparametric specifications
  - Estimation and inference methodology
- Extensions of the Model of Inefficiency
  - Estimation of (in)efficiency
  - Analysis of estimation results
- The Analysis of Panel Data
  - Heterogeneity
  - Technical change
- Statistical Platform for DEA
  - Methodology
  - Reconciliation with SFA
Recent Developments in Econometric Methods

- Model Extensions
- Bayesian Estimation
- Simulation Based Estimation
- Panel Data Methods
- Semiparametric Approaches
- Efficiency Estimation and Inference
- The Interface to DEA
Stochastic Frontier Econometric Model for Inefficiency

\[ y_i = x_i' \beta + \nu_i - u_i \]

Aigner, Lovell, Schmidt (1977)
Meeusen, van den Broeck (1977)
The Econometric Approach to Efficiency Estimation

\[ y_i = x_i'\beta + v_i - u_i \]

\[ f(v_i, u_i) = \ldots \]

\[ \hat{u}_i = f(\text{observable information}) \]

Jondrow et al., Schmidt, Sickles, ... and hundreds of researchers (many of whom are in this room....), 1977 – 2007...
The Normal-Half Normal Model

\[ \ln y_i = \alpha + x_i' \beta + \nu_i - u_i \]

\[ \nu_i \sim N[0, \sigma_v^2] \]

\[ u_i = |U_i|, \quad U_i \sim N[0, \sigma_u^2] \]
The Standard Form

\[
f_z(\varepsilon_i) = \frac{2}{\sqrt{2\pi(\sigma_u^2 + \sigma_v^2)}} \left[ \Phi \left( \frac{-\varepsilon_i (\sigma_u / \sigma_v)}{\sqrt{\sigma_u^2 + \sigma_v^2}} \right) \right] \exp \left( \frac{-\varepsilon_i^2}{2(\sigma_u^2 + \sigma_v^2)} \right) \]

\[
\sigma^2 = (\sigma_u^2 + \sigma_v^2) \text{ and } \lambda = \sigma_u / \sigma_v. \text{ Then,}
\]

\[
f_z(\varepsilon_i) = \frac{2}{\sigma \sqrt{2\pi}} \phi \left( \frac{\varepsilon_i}{\sigma} \right) \left[ \Phi \left( \frac{-\varepsilon_i \lambda}{\sigma} \right) \right].
\]
The Normal-Truncated Normal Model

\[ v_i \sim N(0, \sigma_v^2) \]
\[ U_i \sim N(\mu, \sigma_u^2), u_i = |U_i| \]

\[
\ln L(\alpha, \beta, \sigma, \lambda, \mu) = -N \left[ \ln \sigma + \frac{1}{2} \ln 2\pi + \ln \Phi(\mu/\sigma_u) \right] + \\
\sum_{i=1}^{N} \left[ -\frac{1}{2} \left( \frac{\varepsilon_i + \mu}{\sigma} \right)^2 + \ln \Phi \left( \frac{\mu}{\sigma\lambda} - \frac{\varepsilon_i \lambda}{\sigma} \right) \right]
\]

where \( \sigma_u = \lambda \sigma / \sqrt{1 + \lambda^2} \).
Does the distribution matter?

- Exponential
- Half normal
- Truncated normal
- Other candidates
- Gamma

What do we mean by “matter?”

- Parameter estimates?
- (In)efficiency estimates?
Bayesian Analysis

- **Methodology**
  - General modeling – SFA platform
  - Extensions to Panel Data

- **Applications**
  - Electricity
  - Sports
  - Fishing
  - Hospital costs
  - Farming
  - And so on ... 40+ applications since 2000
Simulation Based Estimators

- Intractible Integrals
  - Bayesian – MCMC methods
  - Classical – Maximum simulated likelihood
- Normal-Gamma Frontier
- Alternative distributions of $v - u$
- The entire cast of recent Bayesian estimators
- Classical approaches to random parameters models
  - SAS: Proc Mixed
  - SAS, Stata, LIMDEP: Integration by simulation
Technological Change in Parametric Models

\[ y_{it} = x_{it}'\beta + v_{it} - u_{it} \]

(1977) (Exponential (and Half Normal): Cross Section, ALS, MLE)

\[ u_{it} \sim \theta \exp(-\theta u_i) \]

(1980) (Truncated Normal: Cross Section, Stevenson, MLE)

\[ u_{it} = |N(\mu, \sigma^2)| \]


\[ u_{it} = u_i \sim \text{half normal} \]

(1990 - 2000) (Gamma: Cross Section, Greene, MSLE, Tsionas, Huang MCMC)

\[ u_{it} \sim \frac{\theta^P \exp(-\theta u_i^P) u_i^{P-1}}{\Gamma(P)} \]

(2004 - 2008) (Generalized Gamma: Random Effects, Griffin and Steel, MCMC)

\[ u_{it} = u_i \sim \frac{\alpha \theta^P \exp(-\theta u_i^{\alpha P}) u_i^\alpha P}{\Gamma(P)} \]

(1977 - 2008) Plus, Fourier transforms, many non- and semiparametrics, etc.)
Simulation and Latent Variables

- An Unobservable Factor (Management, Quality,...)

\[ y_{it} = f(x_{it}, \tilde{m}_i) + v_{it} \]

\[ y_{it} = g(x_{it}, \beta_i, m_i^*) + v_{it} - u_{it} \]

- Applications: Production, Hospital Cost, Measurement Error Models, ...
Model Extensions – Mixture Models

Latent Class or Finite Mixture Models

\[ f(\varepsilon | \Theta) = \sum_{q=1}^{Q} f(\varepsilon | \text{class} = q) \text{Prob}[\text{class} = q | \Theta] \]

- Non (mixed) normality
- Latent heterogeneity
- Other implications of latent classes?
Heteroscedasticity

- Heteroscedasticity in $v_i = \sigma_v^2(z_i)$

- Heteroscedasticity in $u_i = \sigma_u^2(z_i)$

- Is it only variance heterogeneity? $E[u_i]$ and $E[u_i|v_i-u_i]$ are both functions of $\sigma_v^2(z_i)$ and $\sigma_u^2(z_i)$
Semi- and Nonparametric: Kernel Densities

- Nonparametric estimation of the production frontiers
- Nonparametric estimation of $u_i$
- Average derivative (kernel density) estimation of stochastic frontiers
Fundamental Tool - JLMS

\[
E[u_{it} | \varepsilon_{it}] = \left[ \frac{\sigma \lambda}{1 + \lambda^2} \right] \left[ \tilde{\mu}_{it} + \frac{\phi(\tilde{\mu}_{it})}{\Phi(\tilde{\mu}_{it})} \right], \quad \tilde{\mu}_{it} = \frac{-\lambda \varepsilon_{it}}{\sigma}
\]

This estimates \(E[u_i | v_i - u_i]\), not \(u_i\).

Isn’t that what we mean by “estimate \(u_i\)?”

\(E[u_i | \text{available information about } u_i] \)
Efficiency

For the Normal - Truncated Normal Model

\[
E[\exp(-u_i) \mid \varepsilon_i] = \frac{\Phi[(\mu_i^* / \sigma_i^*) - \sigma_i^*]}{\Phi[(\mu_i^* / \sigma_i^*)]} \exp \left[ -\mu_i^* + \frac{1}{2} \sigma_i^* \right]
\]

where \( \mu_i^* = \tilde{\mu}_i + \mu \sigma_u^2 / \sigma^2 \) and \( \sigma_i^2 = \frac{\sigma_u^2 \sigma_v^2}{\sigma^2} \)

For the normal - half normal model, \( \mu = 0. \)
Where do we put the Z’s?

- Production Frontiers: \( f(\text{inputs, “shift factors”}) \)
  - Airlines: Load factor, route map
  - Railroad costs: Network configuration
  - Hospitals: Market factors

- Observed vs. unobserved heterogeneity?
- Does it matter?
The World Health Report - 2000
Application: Health Care Delivery

- Data: 190 countries, 5 years
- Question: How successful?
  - Relative to “the best achievable”
  - Relative to each other
- Methodology: Treat as a production process with outputs, inputs, covariates, and varying degrees of (in)efficiency
Outputs

Life Expectancy
Health Care Attainment

Primary Inputs

\[ X_{it1} = \text{health expenditure per capita in 1997 ppp$} \]
\[ X_{it2} = \text{average years of schooling (and its square)} \]

All variables specified in logarithms
Covariates observed in 1997 only

\[ Z_{i1} = \text{measure of income inequality} \]
\[ Z_{i2} = \text{W.B. measure of freedom and democracy} \]
\[ Z_{i3} = \text{W.B. measure of government effectiveness} \]
\[ Z_{i4} = \text{Dummy variable for location in tropics} \]
\[ Z_{i5} = \text{Population density} \]
\[ Z_{i6} = \text{Public share of health care expenditures} \]
\[ Z_{i7} = \text{Per capita GDP} \]
\[ Z_{i8} = \text{W.B. region designation} \]

These were observed by WHO but not used in the study
Measured Inefficiency

<table>
<thead>
<tr>
<th>Estimated Inefficiency (In Comparison to Fixed Effects Estimates)</th>
<th>DALE</th>
<th>COMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.971</td>
<td>0.928</td>
</tr>
<tr>
<td>Rnk Corr.</td>
<td>0.915</td>
<td>0.910</td>
</tr>
<tr>
<td>Mean OECD Non All</td>
<td>0.0266 (0.0793)</td>
<td>0.0126 (0.0558)</td>
</tr>
<tr>
<td></td>
<td>0.213 (0.256)</td>
<td>0.1543 (0.214)</td>
</tr>
<tr>
<td></td>
<td>0.183 (0.229)</td>
<td>0.132 (0.189)</td>
</tr>
<tr>
<td>Std.Dev.OECD Non All</td>
<td>0.183 (0.0473)</td>
<td>0.0230 (0.0400)</td>
</tr>
<tr>
<td></td>
<td>0.198 (0.188)</td>
<td>0.124 (0.121)</td>
</tr>
<tr>
<td></td>
<td>0.193 (0.185)</td>
<td>0.125 (0.126)</td>
</tr>
</tbody>
</table>

Note OECD vs. Non-OECD:
Per capita GDP: OECD  18,200
Non-OECD  4,450
Inference for Inefficiency

- Confidence Intervals: Horrace and Schmidt, et. al. – Derived true $f(u_i \mid \varepsilon_i)$, lower & upper bounds.
- Bayesian vs. Classical: Kim and Schmidt (2000)
  - Bootstrapping
Horrace/Schmidt Confidence Bounds for Cost Efficiency

The graph shows the Horace/Schmidt confidence bounds for cost efficiency over a range of FIR M values. The x-axis represents FIR M, and the y-axis shows the values of Ee(-t|e). The red line represents the upper bound, and the grey area shows the confidence interval.
One Step – Two Step

- Second step analysis of estimated inefficiencies: Omitted variable biases?
- Tobit (or truncated regression) analysis of DEA results
- Two step estimation in panel data
  - Methodological questions about two step estimation
  - Two step estimation with time invariant effects
Do the markets notice? Is low $u(i)$ associated with good market performance in the life insurance industry? (JPA, 2004)

Using estimated rankings in regulated industries
What is Greene’s Problem

- Connecting factor demands to the parent cost or production function
- Explaining the source of allocative inefficiency
- Devising a tractable econometric model
- What do we do with the results?
- Whither this strand of literature?
Panel Data Frontier Models

- Invariance is a substantive restriction
- Different models produce very different results
- No evidence yet on why the problems arise
Panel Data Estimation

- Extensions of familiar fixed and random effects
  - Bayesian estimators
  - True fixed and random effects with time varying inefficiency
  - What is the incidental parameters problem in this setting?
- Latent variable models
- Hierarchical models
- Random parameters and latent class models
- Reexamination of the fixed effects model
- Time varying inefficiency and heterogeneity – mixtures and latent classes
Panel Data Pursuits

- What do we mean by “inefficiency?”
- Time varying or time invariant?
- Orthogonality to inputs – GMM estimation/
  Hausman/Taylor, Mundlak approaches)
Non-, Semi- and Parametric Approaches

- Schmidt and Sickles, Cornwell et al. (1984, 1990)

\[ y_{it} = \beta' x_{it} + \alpha_i + \nu_{it} \]

- GMM approaches to time varying inefficiency
- Semiparametric approaches
Systematic (Deterministic) Time Variation

- Battese and Coelli (1992)
  \[ y_{i,t} = \alpha + f(x_{i,t}, \beta) + v_{i,t} - \exp[\eta(t - T)]u_i \]

- Kumbhakar and Orea (2003)
  \[ y_{i,t} = \alpha + f(x_{i,t}, \beta) + v_{i,t} - \exp[\eta(t)]u_i \]

  \[ y_{i,t} = \alpha + f(x_{i,t}, \beta) + v_{i,t} - \exp[\eta(z_i, t)]u_i \]
Embedded Variation Truncation Model

\[ u_{i,t} = | U_{i,t} | \]

\[ E[U_{i,t}] = \mu(z_{i,t}, \delta) \]
Time Varying with Invariant Random Effects – Battese/Coelli

\[ y_{i,t} = \alpha + f(x_{i,t}, \beta) + v_{i,t} - g(z_{i,t}, T) | U_i | \]

- Fully Parametric – Normal-Half Normal
- Maximum Likelihood

Just put the dummy variables in the stochastic frontier model.
Freely Time Varying with Fixed Effects

True Fixed Effects Stochastic Frontier

\[ y_{i,t} = \alpha_i + f(x_{i,t}, \beta) + \nu_{i,t} - u_{i,t} \]

- Fully Parametric – Normal-Half Normal
- “Brute Force” Maximum Likelihood
- \( \alpha_i \) is pure heterogeneity
- Methodologically Appropriate
- Statistically Suspect (Incidental Parameters Problem?)
WHO FE Estimates vs. Heterogeneous RE - DALE

Note OECD
Comparing Country Ranks - DALE
Total Efficiency in Production, DALE

Note OECD
Panel Data Frontier Models

- Fixed vs. random effects is not the main issue.

- Disentangling heterogeneity and inefficiency is the main issue.
Fixed vs. Random Effects in the Same Model
True Fixed Effects

Pitt and Lee Random Effects

Random Constants

True Fixed Effects

Regression Fixed Effects

Regression Fixed Effects

Random Constants

Pitt and Lee Random Effects
Model Extensions – Is it Really Noise?

Correlated Error Components

\[ \varepsilon = \nu - u \]

\[ F(\nu, u) = C[F_\nu(\nu), F_u(u)] \]

\[ C[*,*] = \text{parameterized copula function} \]

- Correlation between placement of the frontier and degree of inefficiency?
- Decomposing cost inefficiency?
Or Is It Inefficiency?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Estimated Technical Efficiency</th>
<th>Copula with Half Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Half Normal</td>
<td>Copula with Half Normal</td>
</tr>
<tr>
<td>1</td>
<td>New Mexico .9705 (.7953)</td>
<td>Carolina P&amp;L .8275 (.9449)</td>
</tr>
<tr>
<td>2</td>
<td>Montana .9648 (.7992)</td>
<td>Central La. P&amp;L .8274 (.9435)</td>
</tr>
<tr>
<td>3</td>
<td>Northeast Util .9590 (.8111)</td>
<td>Mississippi Pwr. .8274 (.9508)</td>
</tr>
<tr>
<td>4</td>
<td>Bangor Hydro .9510 (.8254)</td>
<td>Community P. S. .8273 (.9481)</td>
</tr>
<tr>
<td>5</td>
<td>Central Kansas .9504 (.8250)</td>
<td>Kansas G&amp;E .8273 (.9413)</td>
</tr>
<tr>
<td>119</td>
<td>Columbus &amp; So. O. .7601 (.6072)</td>
<td>Columbus &amp; So. O. .6072 (.7601)</td>
</tr>
<tr>
<td>120</td>
<td>Cal. Pacific .7339 (.4805)</td>
<td>United Gas. Illum. .5233 (.7180)</td>
</tr>
<tr>
<td>121</td>
<td>United Gas Illum. .7180 (.5233)</td>
<td>Cal. Pacific .4805 (.7339)</td>
</tr>
<tr>
<td>122</td>
<td>Northwestern P.S. .7032 (.4766)</td>
<td>Northwestern P.S. .4766 (.7032)</td>
</tr>
<tr>
<td>123</td>
<td>Maine Public Svc. .6786 (.3990)</td>
<td>Maine Public Svc. .3990 (.6786)</td>
</tr>
</tbody>
</table>

Sample Technical Efficiency

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.8877</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.0543</td>
</tr>
<tr>
<td>Min</td>
<td>0.6786</td>
</tr>
<tr>
<td>Max</td>
<td>0.9705</td>
</tr>
</tbody>
</table>

Table from Smith, M., “Stochastic Frontier Models with Dependent Error Components,” WP, University of Sydney (Also Bandyopadhyay & Das in JPA, 2006)
How Good Is JLMS?

- Estimator of $u_i$:
  $$\hat{u}_i = \hat{E}(u_i | \hat{\theta}, e_i)$$

- Confidence intervals based on true distribution

- $\text{plim} \; \hat{u}_i - u_i \neq 0$ (Variance $\not\to 0$)

- Behavior of the JLMS Estimator

- How good is the conditional mean as the estimator?

- And now, ... is it really free of the noise?
Efficiency Estimation

“How much confidence should we place in efficiency estimates?”

A. Street, Health Economics

“relative (...) efficiency are sensitive to estimation decisions and (...) little confidence can be placed in the point estimates for individual (...)”
The Interface Between SFA and DEA

- What is the DGP?
- If there is a DGP, what does DEA estimate?
- DEA is biased? For what? Is SFA unbiased?
- Consistency
  - What do we mean by consistency?
  - Is DEA or SFA consistent?
Statistical Issues in DEA

- Leopold Simar, Paul Wilson et al. extensions of econometrics to analysis of DEA scores. (JPA, 2000) + several dozen other papers and authors

- The Incidental Parameters Problem: “The Curse of Dimensionality”

- Are the estimates (DEA, SFA) really similar? A cautionary note: Beware of correlations.
An Experiment

- The Christensen and Greene Data (again)
  - Relatively clean
  - Standard test data set (esp. Bayesian studies)
- 123 firms, cross section, 1970
- Production: Output in Millions of KWH
- Inputs: Capital, Labor and Fuel
Stochastic Frontier

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>LOGQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variances: Sigma-squared(v)=</td>
<td>.00946</td>
</tr>
<tr>
<td>Sigma-squared(u)=</td>
<td>.05114</td>
</tr>
<tr>
<td>Sigma(v) =</td>
<td>.09727</td>
</tr>
<tr>
<td>Sigma(u) =</td>
<td>.22615</td>
</tr>
<tr>
<td>Sigma = Sqr[(s^2(u)+s^2(v)]=</td>
<td>.24618</td>
</tr>
<tr>
<td>Stochastic Production Frontier, e=v-u.</td>
<td></td>
</tr>
</tbody>
</table>

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|----------|---------|-----------|
| Constant | 8.38161612  | .27481432      | 30.499   | .0000   |           |
| LOGFUEL  | 1.10044551  | .04652336      | 23.654   | .0000   | -.86174332|
| LOGLABOR | -.17015348  | .04142942      | -4.107   | .0000   | -7.98062905|
| LOGCAP   | .15807939   | .05010930      | 3.155    | .0016   | -2.78214328|

<p>| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|&gt;z] | Mean of X |
|----------|-------------|----------------|----------|---------|-----------|
| Lambda   | 2.32505529  | .38676618      | 6.012    | .0000   |           |
| Sigma    | .24617735   | .00143743      | 171.262  | .0000   |           |</p>
<table>
<thead>
<tr>
<th>Data Envelopment Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output Variables: Q</td>
</tr>
<tr>
<td>Input Variables: FUEL LABOR CAPITAL</td>
</tr>
<tr>
<td>Underlying Technology assumes VARIABLE Returns to Scale.</td>
</tr>
<tr>
<td>Estimated Efficiencies: Mean Std.Deviation Minimum Maximum</td>
</tr>
<tr>
<td>Technical Efficiency Input Oriented .7692 .1390 .3464 1.0000</td>
</tr>
<tr>
<td>Output Oriented .7657 .1467 .2960 1.0000</td>
</tr>
<tr>
<td>Sample Size: 123 Observations. 123 Complete observations</td>
</tr>
<tr>
<td>Descriptive Statistics for SF Model</td>
</tr>
<tr>
<td>Mean = .850663</td>
</tr>
<tr>
<td>Standard Deviation = .077037</td>
</tr>
<tr>
<td>Minimum = .590270</td>
</tr>
<tr>
<td>Maximum = .974458</td>
</tr>
</tbody>
</table>
Simulate Output Using SF Estimates of RHS Parameters and Var[u], Var[v]

+---------------------------------------------+
| Dependent variable                  LYS     |
| Variances: Sigma-squared(v)=       .00397   |
|            Sigma-squared(u)=       .06171   |
|            Sigma(v)        =       .06304   | (.09727)
|            Sigma(u)        =       .24841   | (.22615)
| Sigma = Sqr[(s^2(u)+s^2(v)]=       .25628   |
| Stochastic Production Frontier, e=v-u.      |
+---------------------------------------------+

+--------+--------------+----------------+--------+--------+----------+
|Variable| Coefficient  | Standard Error |b/St.Er.|P[|Z|>z]| Mean of X|
|--------+--------------+----------------+--------+--------+----------+
|---------|Primary Index Equation for Model
|Constant|    8.41620291       .22113669    38.059   .0000
|LOGFUEL |    1.09675159       .03682529    29.783   .0000   -.86174332
|LOGLABOR|    -.17717129       .03295127    -5.377   .0000  -7.98062905
|LOGCAP  |     .17769642       .03708025     4.792   .0000  -2.78214328
|---------|Variance parameters for compound error
|Lambda  |    3.94019897      1.36276324     2.891   .0038
|Sigma   |     .25628384       .00171499   149.438   .0000
|
Estimated u(i) Predicting Actual u(i)
DEA Predicting Actual Known u(i)
DEA vs. SFA Based on Known Actual u(i)
Stochastic Frontier Analysis

Data Envelopment Analysis