Aalto University
School of Economics

How Operational Conditions and Practices Affect Productive Performance?

Efficient Semi-parametric One-Stage Estimators

Andrew Johnson (Texas A&M) and Timo Kuosmanen (Aalto)

Kansantaloustieteen päivät
4-5 Feb 2010
Tampere

Example: efficiency analysis of railroads

Outputs (y)

- Passenger service (total passenger-km)
- Freight service (total tonne-km)

Inputs (x)

- Total train mileage (total train-km)
- Train capacity (gross train tonne-km)

Contextual variables (z)

- Geographic factors (population. land area, region)
- Competition (market shares in passenger and freight services)
- Government regulation (ownership of railway track)

Model

$$y = f(\mathbf{x}) \cdot e^{\mathbf{z}\boldsymbol{\beta} - u + v}$$

where

- *f* is a nonparametric. monotononic increasing and concave production function
- **β** is a parameter vector representing marginal effects of contextual variables on efficiency
- *u* is a random non-negative inefficiency term
- *v* is a random symmetric noise term

Efficiency interpretation

$$\frac{y}{f(\mathbf{x})} = e^{\mathbf{z\beta} - u + v}$$

where

- $y / f(\mathbf{x})$ is output efficiency
- $\boldsymbol{\beta}$ is a parameter vector representing marginal effects of contextual variables on efficiency
- u is a random non-negative inefficiency term
- v is a random symmetric noise term

Log-transformed model

$$\ln y = \ln f(\mathbf{x}) + \mathbf{z}\boldsymbol{\beta} + v - u$$

where

- $\boldsymbol{\beta}$ is a parameter vector representing marginal effects of contextual variables on efficiency
- u is a random non-negative inefficiency term
- v is a random symmetric noise term

Semiparametric partial linear model

$$y = f(\mathbf{x}) + \mathbf{z}\boldsymbol{\beta} + \varepsilon$$

Robinson (1988), Newey (1994), Linton (1995) in *Econometrica*

- Nonparametric part: $f(\mathbf{x})$
- Parametric part: $\mathbf{z}\boldsymbol{\beta}$
- Disturbance term: ε

2-stage DEA estimation (Ray 1988)

Stage 1:

Estimate efficiency by data envelopment analysis DEA

$$y / f(\mathbf{x})$$

Stage 2:

Use the logarithm of the DEA efficiency estimate as the dependent variable of the regression

$$\ln[y / f(\mathbf{x})] = \mathbf{z}\boldsymbol{\beta} + e$$

2-stage DEA estimation

Banker and Natarajan (2008). *Oper. Res.*

- 2-DEA estimator of coefficients **β** is consistent
- Monte Carlo simulations: 2-DEA performs better than parametric regression where *f* misspecified
- Restrictive assumptions:
  - Inputs **x** independent of contextual variables **z**
  - Effect of context is one-sided: $\mathbf{z} \le \mathbf{0}$
  - Truncated noise $|v| \le V^M$

2-stage DEA estimation

Johnson and Kuosmanen (2009) WP

- 2-DEA estimator of coefficients **β** is consistent
- Assumptions relaxed:
  - Inputs **x** not perfectly dependent of contextual variables **z**
  - Effect of context is bounded from above: $\mathbf{z}\boldsymbol{\beta} \leq \boldsymbol{\zeta}$
  - Truncated noise $|v| < V^M$

- In finite samples, 2-DEA is biased:

$$Bias(\boldsymbol{\beta}) = (\mathbf{Z'Z})^{-1}\mathbf{Z'}Bias(f^{DEA}(\mathbf{X}))$$

2-stage DEA estimation

Simar and Wilson (2007 *J. Ectr.*) critique:

- DEA estimator does not converge fast enough
- Inferences on coefficients **β** (significance tests. conf. Intervals) are invalid. Bootstrap method proposed.
- Restrictive assumptions:
  - Inefficiency u has truncated normal distribution
  - No noise: $v=0$

Efficient 1-stage CNLS estimator

Strategy: Estimate equation

$$\ln y = \ln f(\mathbf{x}) + \mathbf{z}\boldsymbol{\beta} + \varepsilon$$

by convex nonparametric least squares (CNLS) (Kuosmanen. 2008. *Etcr. J*)

Firm-specific inefficiencies *u* can be estimated from residuals *e*:

- In cross-section: conditional expectation *E*[*u*|*e*] (Jondrow et al. 1982)
- In panel data: fixed or random effects

Efficient 1-stage CNLS estimator

CNLS problem with **z** variables

$$\min_{f,\boldsymbol{\beta},\varepsilon} \sum_{i=1}^{n} \varepsilon_i^2$$

$$s.t.$$

$$\ln y_i = \ln f(\mathbf{x}_i) + \mathbf{z}_i \boldsymbol{\beta} + \varepsilon_i \quad \forall i$$

f is monotonic increasing and concave

Efficient 1-stage CNLS estimator

Kuosmanen (2008): problem can be equivalently stated as

$$\min_{\phi,\alpha,\boldsymbol{\gamma},\boldsymbol{\beta},\varepsilon} \sum_{i=1}^{n} \varepsilon_i^2$$

$$s.t.$$

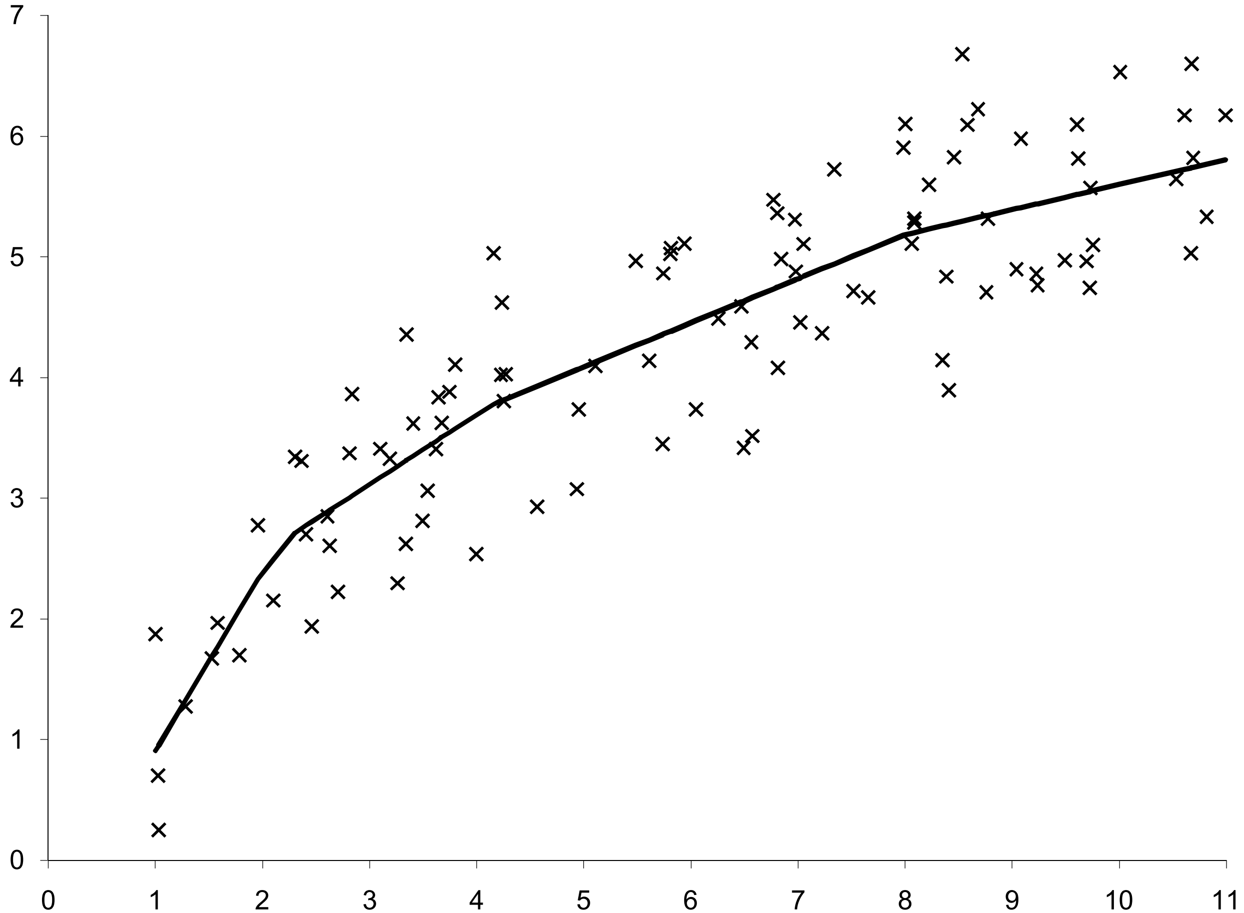
$$\ln y_i = \ln \phi_i + \mathbf{z}_i \boldsymbol{\beta} + \varepsilon_i \quad \forall i$$

$$\phi_i = \alpha_i + \mathbf{x}_i \boldsymbol{\gamma}_i \quad \forall i$$

$$\alpha_i + \mathbf{x}_i \boldsymbol{\gamma}_i \le \alpha_h + \mathbf{x}_i \boldsymbol{\gamma}_h \quad \forall i, h$$

$$\boldsymbol{\gamma}_i \ge \mathbf{0} \quad \forall i$$

Illustration of CNLS curve



Efficient 1-stage CNLS estimator

Johnson & Kuosmanen (2009) WP

Statistical properties of CNLS estimator of **β**

- Consistent
- Asymptotically normal
- Asymptotically efficient
- Converges at standard parametric rate $n^{-1/2}$

=> Conventional methods of statistical inference (*t*-tests, confidence intervals, etc.) apply

Monte Carlo simulations

Banker and Natarajan (2008) simulations replicated

Root Mean Squared Deviation (RMSD) at different levels of correlation (ρ) between x and z variables

Estimator	ρ=-0.8	ρ=-0.2	**ρ= 0**	ρ=+0.2	ρ=+0.8
OLS: y on z			**117**		
2-DEA	74.0	22.0	**18.9**	20.6	57.6
CNLS	29.1	18.3	**17.9**	18.3	30.1

Application to railroads

- Cross-section of 73 railroad firms
- Year 2008
- Data source: http://www.uic.org

- Joint production: passenger and freight services
- Efficiency modeled by using Shephard's output distance function

Application to railroads

Outputs (y)

- Passenger service (total passenger-km)
- Freight service (total tonne-km)

Inputs (x)

- Total train mileage (total train-km)
- Train capacity (gross train tonne-km)

Contextual variables (z)

- Geographic factors (population, land area, region)
- Competition (market shares in passenger and freight services)
- Government regulation (ownership of railway track)

Application to railroads

Preliminary results: area dummies (EU = 0)

Area	**Coefficient**	**Std. error**	**P value**
America	-64.6	35.6	0.074
EFTA	-55.7	23.1	0.019
Asia	85.6	19.3	0.000
CEEC	107.9	19.9	0.000
Africa	138.1	18.7	0.000
CIS	146.3	24.1	0.000

Application to railroads

Preliminary results: operational conditions

Variable	Coeff.	Std. Error	P value
land area	0.071	0.003	0.000
population	-4.205	0.030	0.000
market share in passengers	-143.0	15.37	0.000
market share in freight	-72.81	15.00	0.000
track ownership	18.15	12.89	0.164

Conclusions

- Joint estimation of nonparametric frontier and effects of contextual variables possible by CNLS
- CNLS allows for unbounded noise and correlation between inputs and contextual variables
- CNLS estimator of contextual variables is consistent, asymptotically efficient, asymptotically normal, and converges at standard parametric rate
- CNLS superior to 2-DEA in Monte Carlo simulations
- Using conventional standard errors, t-tests, p-values ,… convenient in applications

Thank you for your attention!

Internet: http://www.nomepre.net/stoned/

E-mail: timo.kuosmanen@hse.fi