Dealing with the endogeneity issue in the estimation of educational efficiency using DEA

Daniel Santín
Gabriela Sicilia
Complutense University of Madrid

Efficiency in Education Workshop
19th-20th September 2014
London, UK

Outline

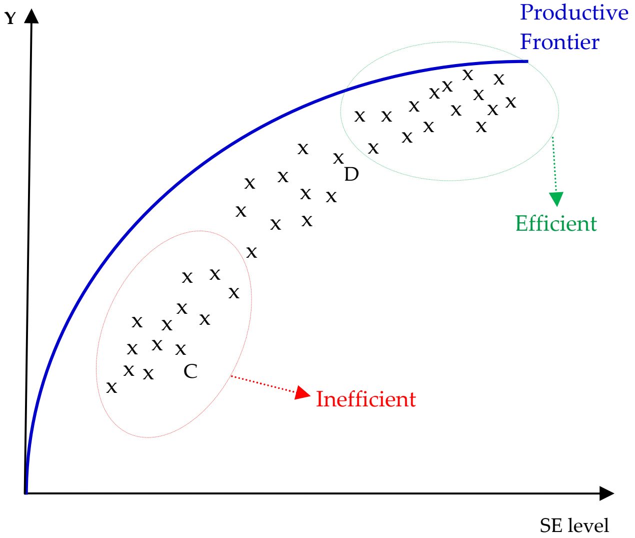
Endogeneity in Education - Self-selection

- Endogeneity is one of the most important concerns in Education Economics (Schottler et al. 2011)
- Better schools attract relatively more advantaged students (high socio-economic level and more motivated parents)
- Parent motivation (unobserved) is positively correlated with SEL.
- These pupils (and thus the school they attend) will tend to obtain better academic results for two reasons:
  1. $\uparrow$ SEL which is an essential input
  2. $\uparrow$ Motivated students which are more efficient

Positive correlation between the input and school efficiency

Schools with students from a high SEL are more prone to be efficient

Endogenous input in a single-input single-output set



The endogeneity issue in non-parametric techniques

- Endogeneity was widely studied in the econometrics, but little in non-parametric frontier techniques (Gong and Sickles 1992, Orme and Smith 1996, Bifulco and Bretschneider 2001, Ruggiero 2004)
- *A priori* it seems that this problem does not affect DEA estimates, since no assumptions about parametric functional form
- But, as Kuosmanen and Johnson (2010) demonstrate that DEA can be formulated as a non-parametric least-squares model under the assumption that $\epsilon_i \leq 0$
- If $E(\epsilon|X) \neq 0$, then efficiency estimates $(\hat{\varphi}_i)$ can be biased
- In a recent work Cordero et al. (2013) show using MC that although DEA is robust to negative endogeneity, a significant positive correlation severely biases DEA performance

How can be DEA estimates be affected when $E(\varphi|X) \neq 0$?

	Spearman's correlation	MAE	% Assigned two or more quintiles from actual	% Correctly assigned to bottom quintile	% Assigned to bottom quintile actually in the two first quintiles	% Assigned to top quintile actually in the two last quintiles
ρ = 0.0	**0.73**	**0.07**	**13.4**	**74.7**	**0.1**	**11.2**
ρ = 0.8	**0.27**	**0.12**	**38.4**	**34.2**	**12.6**	**34.2**
ρ = 0.4	**0.59**	**0.09**	**20.7**	**62.7**	**0.9**	**62.7**

Note: Mean values after 1,000 replications. Sample size N=100. Translog DGP. DEA estimated under VRS

Source: Cordero, JM.; Santín, D. and Sicilia, G. "*Dealing with the Endogeneity Problem in Data Envelopment Analysis*", MPRA, April 2013.

How to deal with this problem?

1. How can we identify the presence of an endogenous input in an empirical research?

2. How can we deal with this issue in order to improve DEA estimations?

How to identify this problem?

A simple procedure for detecting the presence of positive endogenous inputs in empirical applications:

1. From the empirical dataset $\chi = \{(X_i, Y_i)\ i = 1, ..., n\}$ randomly draw with replacement a bootstrap sample $\chi_b^* = \{(X_{ib}^*, Y_{ib}^*)\ i = 1, ..., n\}$
2. Estimate $\hat{\theta_{ib}^*}\ \ i = 1, ..., n$ using DEA LP
3. For each input $k = 1, ..., p$ compute $\rho_{kb}^* = corr(x_{ik}^*, \hat{\theta_i^*})\ \ i = 1, ..., n$
4. Repeat steps 1-3 B times in order to obtain for $k = 1, ..., p$ a set of correlations: $\{\rho_{kb}^*,\ \ b = 1, ..., B\}$

How to identify this problem?

5. Compute $\gamma_k^* = \dfrac{1}{B}\sum_{b=1}^{B}[I_{[0,1]}(\rho_k^*)]_b \;\; for \;\; k = 1, ..., p$

   where $I_{[0,1]}(\rho_k^*)$ is the Indicator Function defined by:

$$I_{[0,1]}(\rho_k^*) = \begin{cases} 1, & \text{if } 0 \leq \rho_k^* \leq 1; \\ 0, & \text{otherwise.} \end{cases}$$

6. Finally, classify each input using the following criterion:
   - If $\gamma_k^* < 0.25 \rightarrow$ Exogenous/Negative endogenous input *k*
   - If $0.25 \leq \gamma_k^* < 0.5 \rightarrow$ Positive LOW endogenous input *k*
   - If $0.5 \leq \gamma_k^* < 0.75 \rightarrow$ Positive MIDDLE endogenous input *k*
   - If $\gamma_k^* \geq 0.75 \rightarrow$ Positive HIGH endogenous input *k*

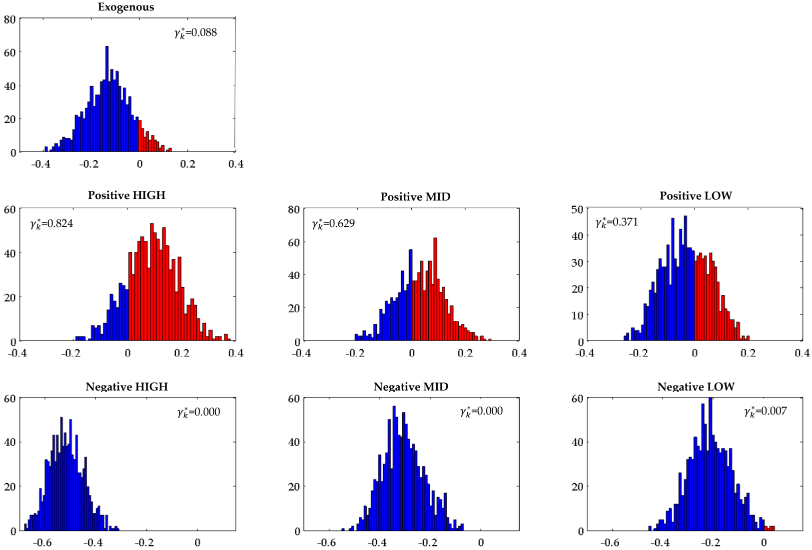
We propose to combine the IV approach (*e.g.*,Greene, 2003) with DEA model by instrumenting the endogenous input.

1. Find an instrumental input(Z) that satisfies:
    - Is correlated with the endogenous input(x_e), *i.e.* $E(x_e|Z) \neq 0$
    - Is exogenous from true efficiency, *i.e.* $E(\epsilon|Z) = 0$
2. Isolate the part of (x_e) that is uncorrelated with the efficiency by regressing $x_{ei} = \alpha + \beta_1 x_1 i + ... + \beta_k x_{ki} + \delta Z_i + \xi_i$ and computing $\hat{x}_{ei}$
3. Replace the endogenous input (x_e) by $\hat{x}_{ei}$ and estimate DEA efficiency scores for each DMU ($\hat{\varphi}_i$)

MC experimental design

- Single-output multi-input framework. We follow the same simple DGP as in CSS (2013) to compute, Y, X, u, and v.
- True efficiency (u_i) is exogenous from x_1 and x_2.
- Seven different scenarios with different levels of correlations between u_i and x_3 $\rho = \{-0.8, -0.4, -0.2, 0, 0.2, 0, 4, 0.8\}$.
- We generate Z$\sim U[5, 50]$ uncorrelated with true efficiency $E(u|Z) = 0$ and moderately correlated with the endogenous input x_3, where $E(x_3|Z) \simeq 0.25$
- Cobb-Douglas and Translog DGP, N={40,100,400}, and B=1,000
- We compare estimations from the conventional DEA and from II-DEA.

MC results - Input classification criterio



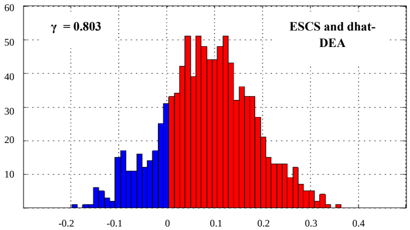
MC results - II-DEA Accuracy measures

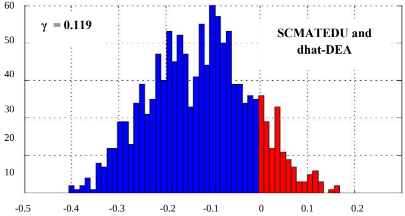
		Spearman's correlation	MAE	% Assigned two or more quintiles from actual	% Correctly assigned to bottom quintile	% Assigned to bottom quintile actually in the two first quintiles	% Assigned to top quintile actually in the two last quintiles
ρ = 0.0	**DEA**	**0.73**	**0.072**	**13.3**	**74.8**	**0.2**	**12.3**
ρ = 0.8	**DEA**	**0.34**	**0.116**	**34.8**	**40.8**	**8.2**	**30.3**
	II-DEA	**0.76**	**0.097**	**10.0**	**75.7**	**0.1**	**15.6**
ρ = 0.4	**DEA**	**0.61**	**0.085**	**19.8**	**64.8**	**0.7**	**18.6**
	II-DEA	**0.66**	**0.099**	**17.1**	**62.6**	**4.0**	**16.8**

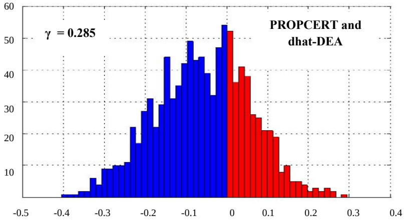
Note: Mean values after 1,000 replications. Sample size N=100. Translog DGP. DEA estimated under VRS

- Highly stratified Uruguayan education system (strong correlation between SEL and academic results)
- Data from PISA 2012, N = 71, p = 3, q = 1.
- Output (y): result in mathematics (maths)
- Inputs (X):
  - School Quality Educational Resources Index (SCMATEDU)
  - Proportion of Certified Teachers (PROPCERT)
  - **Socio-economic Level Index (ESCS)** - potential endogenous input
- Instrumental input (Z): "Pct. of students who access to Internet before thirteen" (ACCINT); where $\rho_{(ESCS,ACCINT)} = 0.20$

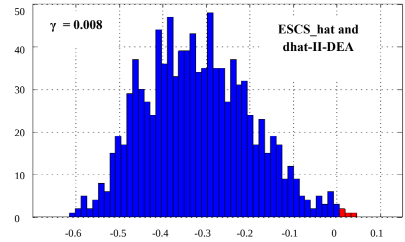
Detection criteria for ESCS in Uruguayan public secondary schools

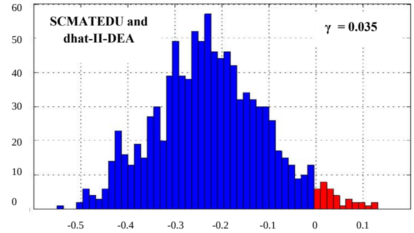


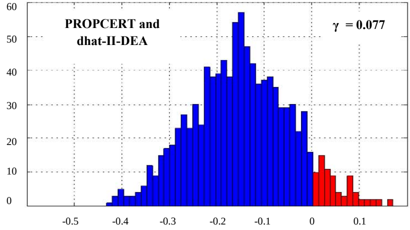




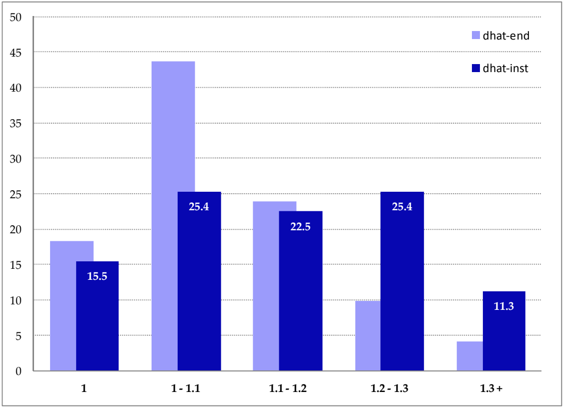
Detection criteria for ESCS-hat in Uruguayan public secondary schools







II-DEA estimates



Efficiency	Mean	Std-Dev.	Min.	Max.
dhat-end	1.101	0.102	1.000	1.468
dhat-inst	1.167	0.149	1.000	1.640

Quintiles by ESCS	Mean ESCS	Mean dhat-inst	Mean dhat-end	Mean |Bias|
Bottom quintile	**1.68**	1.286	1.079	0.206
4th quintile	**1.92**	1.229	1.132	0.097
3rd quintile	**2.13**	1.146	1.107	0.050
2nd quintile	**2.40**	1.106	1.108	0.011
Top quintile	**2.82**	1.076	1.079	0.003

Source: Author's estimates using PISA 2012 data

Semi-parametric two-stage model results

Dependent variable: dhat	Truncated + bootstrap (II-DEA)				Truncated + bootstrap (DEA)			
	Coef	**Std. Err.**	**z**		**Coef**	**Std. Err.**	**z**	
TECHVOC[a]	0.0097	0.057	0.17		0.0536	0.990	0.32	
RURAL[a]	-0.0062	0.074	-0.08		-0.0255	0.087	-0.29	
SCHSIZE	-0.0001	0.000	-1.81	*	-0.0001	0.000	-1.53	
PCTGIRL	0.0249	0.165	0.15		-0.1433	0.166	-0.87	
ICTSCH	-0.0395	0.067	-0.59		-0.0395	0.049	-0.80	
PCTCORRECT	-0.2898	0.117	-2.47	**	-0.1300	0.089	-1.46	
ANXMAT	0.2410	0.077	3.14	***	0.1255	0.064	1.96	**
PCTMATHEART	0.5081	0.268	1.89	*	-0.0087	0.243	-0.04	
TEACHGOAL	0.3965	0.253	1.57		-0.3214	0.227	-1.41	
TEACHCHECK	-0.5443	0.228	-2.39	**	-0.0017	0.189	-0.01	
HINDTEACH[a]	-0.0873	0.039	-2.24	**	-0.0497	0.037	-1.35	
TEACHMORAL[a]	-0.1056	0.049	-2.13	**	-0.0253	0.036	-0.71	
RESPCUR	-0.0962	0.064	-1.50		-0.0661	0.072	-0.92	
RESPRES	0.1902	0.199	0.95		0.1696	0.221	0.77	
_cons	0.5361	0.423	1.27		1.0170	0.401	2.53	
/sigma	0.0926	0.01	8.65		0.0751	---	---	

Note: 'Coef' is the estimated coefficient, S.E. is the robust standard error of the coefficient estimate.
N = 71. ***p-value < 0.01 ; **p-value < 0.05 ; *p - value < 0.10
Source: Author's estimations using PISA 2012 data.

Concluding remarks

- We propose a simple and effective criterion to **detect endogenous inputs** in DEA empirical applications
- MC experiments also suggest that the proposed strategy **II-DEA outperforms conventional DEA** when ρ is significantly high positive.
- Taking into account the presence of high positive endogeneity has **major implications in educational policy recommendations**
- More research is needed:
  - Derive the asymptotic properties of the II-DEA estimator
  - Adapt to our context some previous proposed testing procedures for independence (*e.g.*Peyrache and Coelli 2009)
  - Extend the analysis to multi-output sets

Thanks...!

Daniel Santín
(dsantin@ccee.ucm.es)

Gabriela Sicilia
(gabriels@ucm.com)

Dealing with the endogeneity issue in the estimation of educational efficiency using DEA

Daniel Santín
Gabriela Sicilia
Complutense University of Madrid

Efficiency in Education Workshop
19th-20th September 2014
London, UK