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

1. The endogeneity issue
2. How to identify this problem?
3. How to deal with it?
4. Monte Carlo simulations
5. Empirical application
6. Concluding remarks
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. ↑ SEL which is an essential input
2. ↑ 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

Figure 2. True frontier and DEA-BCC estimates in a positive endogenous scenario
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 (\( \hat{\phi}_i \)) 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 $E(\varphi|X) \neq 0$?

<table>
<thead>
<tr>
<th>Spearman's correlation</th>
<th>MAE</th>
<th>% Assigned two or more quintiles from actual</th>
<th>% Correctly assigned to bottom quintile</th>
<th>% Assigned to bottom quintile actually in the two first quintiles</th>
<th>% Assigned to top quintile actually in the two last quintiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho = 0.0 )</td>
<td>0.73</td>
<td>0.07</td>
<td>13.4</td>
<td>74.7</td>
<td>0.1</td>
</tr>
<tr>
<td>( \rho = 0.8 )</td>
<td>0.27</td>
<td>0.12</td>
<td>38.4</td>
<td>34.2</td>
<td>12.6</td>
</tr>
<tr>
<td>( \rho = 0.4 )</td>
<td>0.59</td>
<td>0.09</td>
<td>20.7</td>
<td>62.7</td>
<td>0.9</td>
</tr>
</tbody>
</table>

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 \( \chi = \{(X_i, Y_i) \mid i = 1, \ldots, n\} \) randomly draw with replacement a bootstrap sample \( \chi^*_b = \{(X^*_ib, Y^*_ib) \mid i = 1, \ldots, n\} \)

2. Estimate \( \hat{\theta}^*_ib \mid i = 1, \ldots, n \) using DEA LP

3. For each input \( k = 1, \ldots, p \) compute \( \rho^*_kb = corr(x^*_ik, \hat{\theta}^*_i) \mid i = 1, \ldots, n \)

4. Repeat steps 1-3 B times in order to obtain for \( k = 1, \ldots, p \) a set of correlations: \( \{\rho^*_kb, \ b = 1, \ldots, B\} \)
How to identify this problem?

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

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 \( \gamma_k^* < 0.25 \) → Exogenous/Negative endogenous input \( k \)
- If \( 0.25 \leq \gamma_k^* < 0.5 \) → Positive LOW endogenous input \( k \)
- If \( 0.5 \leq \gamma_k^* < 0.75 \) → Positive MIDDLE endogenous input \( k \)
- If \( \gamma_k^* \geq 0.75 \) → Positive HIGH endogenous input \( k \)
How to deal with endogeneity in DEA applications?

The “Instrumental Input” DEA propose (II-DEA)

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 \( x_{ei} = \alpha + \beta_1 x_1 i + \ldots + \beta_k x_k i + \delta Z_i + \xi_i \) and computing \( \hat{x}_{ei} \)

3. Replace the endogenous input \( (x_e) \) by \( \hat{x}_{ei} \) and estimate DEA efficiency scores for each DMU (\( \hat{\phi}_i \))
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 $E(x_3|Z) \approx 0.25$.

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 criteria

\( \gamma_k^* = 0.088 \)

\( \gamma_k^* = 0.824 \)

\( \gamma_k^* = 0.629 \)

\( \gamma_k^* = 0.371 \)

\( \gamma_k^* = 0.007 \)

\( \gamma_k^* = 0.000 \)

\( \gamma_k^* = 0.007 \)
### MC results - II-DEA Accuracy measures

<table>
<thead>
<tr>
<th></th>
<th>Spearman's correlation</th>
<th>MAE</th>
<th>% Assigned two or more quintiles from actual</th>
<th>% Correctly assigned to bottom quintile</th>
<th>% Assigned to bottom quintile actually in the two first quintiles</th>
<th>% Assigned to top quintile actually in the two last quintiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0.0$</td>
<td>DEA</td>
<td>0.73</td>
<td>0.072</td>
<td>13.3</td>
<td>74.8</td>
<td>0.2</td>
</tr>
<tr>
<td>$\rho = 0.8$</td>
<td>DEA</td>
<td>0.34</td>
<td>0.116</td>
<td>34.8</td>
<td>40.8</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>II-DEA</td>
<td>0.76</td>
<td>0.097</td>
<td>10.0</td>
<td>75.7</td>
<td>0.1</td>
</tr>
<tr>
<td>$\rho = 0.4$</td>
<td>DEA</td>
<td>0.61</td>
<td>0.085</td>
<td>19.8</td>
<td>64.8</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>II-DEA</td>
<td>0.66</td>
<td>0.099</td>
<td>17.1</td>
<td>62.6</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Note: Mean values after 1,000 replications. Sample size N=100. Translog DGP. DEA estimated under VRS.
Empirical application
The Uruguayan public secondary schools

- 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

- ESCS and dhat-DEA
  - $\gamma = 0.803$

- SCMATEDU and dhat-DEA
  - $\gamma = 0.119$

- PROPCERT and dhat-DEA
  - $\gamma = 0.285$
Detection criteria for ESCS-hat in Uruguayan public secondary schools

\[ \gamma = 0.008 \]

ESCS_hat and dhat-II-DEA

\[ \gamma = 0.035 \]

SCMATEDU and dhat-II-DEA

\[ \gamma = 0.077 \]

PROPCERT and dhat-II-DEA

Santín, D. and Sicilia, G. ( )
Dealing with endogeneity...
II-DEA estimates

<table>
<thead>
<tr>
<th></th>
<th>Efficiency Mean</th>
<th>Std-Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>dhat-end</td>
<td>1.101</td>
<td>0.102</td>
<td>1.000</td>
<td>1.468</td>
</tr>
<tr>
<td>dhat-inst</td>
<td>1.167</td>
<td>0.149</td>
<td>1.000</td>
<td>1.640</td>
</tr>
</tbody>
</table>

Quintiles by ESCS

<table>
<thead>
<tr>
<th>Quintile</th>
<th>Mean ESCS</th>
<th>Mean dhat-inst</th>
<th>Mean dhat-end</th>
<th>Mean</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom quintile</td>
<td>1.68</td>
<td>1.286</td>
<td>1.079</td>
<td>0.206</td>
<td></td>
</tr>
<tr>
<td>4th quintile</td>
<td>1.92</td>
<td>1.229</td>
<td>1.132</td>
<td>0.097</td>
<td></td>
</tr>
<tr>
<td>3rd quintile</td>
<td>2.13</td>
<td>1.146</td>
<td>1.107</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>2nd quintile</td>
<td>2.40</td>
<td>1.106</td>
<td>1.108</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Top quintile</td>
<td>2.82</td>
<td>1.076</td>
<td>1.079</td>
<td>0.003</td>
<td></td>
</tr>
</tbody>
</table>

Source: Author’s estimates using PISA 2012 data
Semi-parametric two-stage model results

<table>
<thead>
<tr>
<th>Dependent variable: dhat</th>
<th>Truncated + bootstrap (II-DEA)</th>
<th>Truncated + bootstrap (DEA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>TECHVOC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.0097</td>
<td>0.057</td>
</tr>
<tr>
<td>RURAL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-0.0062</td>
<td>0.074</td>
</tr>
<tr>
<td>SCHSIZE</td>
<td>-0.0001</td>
<td>0.000</td>
</tr>
<tr>
<td>PCTGIRL</td>
<td>0.0249</td>
<td>0.165</td>
</tr>
<tr>
<td>ICTSCH</td>
<td>-0.0395</td>
<td>0.067</td>
</tr>
<tr>
<td>PCTCORRECT</td>
<td>-0.2898</td>
<td>0.117</td>
</tr>
<tr>
<td>ANXMAT</td>
<td>0.2410</td>
<td>0.077</td>
</tr>
<tr>
<td>PCTMATHEART</td>
<td>0.5081</td>
<td>0.268</td>
</tr>
<tr>
<td>TEACHGOAL</td>
<td>0.3965</td>
<td>0.253</td>
</tr>
<tr>
<td>TEACHCHECK</td>
<td>-0.5443</td>
<td>0.228</td>
</tr>
<tr>
<td>HINDTEACH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-0.0873</td>
<td>0.039</td>
</tr>
<tr>
<td>TEACHMORAL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-0.1056</td>
<td>0.049</td>
</tr>
<tr>
<td>RESPCUR</td>
<td>-0.0962</td>
<td>0.064</td>
</tr>
<tr>
<td>RESPRES</td>
<td>0.1902</td>
<td>0.199</td>
</tr>
<tr>
<td>_cons</td>
<td>0.5361</td>
<td>0.423</td>
</tr>
<tr>
<td>/sigma</td>
<td>0.0926</td>
<td>0.01</td>
</tr>
</tbody>
</table>

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 $\rho$ 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