Propensity Scoring Approaches to Address Problems with Two-Stage DEA and Stochastic Frontier Analyses

On a Set of Methods to Address Problems with Two-Stage DEA

Pavlo Demchuk and Robin Sickles

Efficiency Measurement: New Methods and Application to the Food Sector
Toulouse, June 27-28, 2011

Overview

- Motivation
- 2 Stage DEA
- Propensity Score Analysis (PSA)
- PSA Matching
- Simulation
- Application: Texas Electricity Market
- Conclusions

DEA Setup (Two Stage)

Stage 1

- N decision making units (DMUs)
- Linear function of n inputs that produce m outputs
- There is no explicit production or cost function
- Linear function of inputs x is minimized given the linear function of outputs y (linear function of outputs is maximized given the linear function of inputs)
- Efficiency scores

Stage 2

- Assume a relationship between efficiency scores and some explanatory variables z (done using Truncated Reg with Bootstrap)

Motivation

- Bias due to the misspecification of the model in stage 1 (omitted variable)
- Selection bias
- Group Comparison

Propensity Score Analysis

- One analyzes causal effects of treatment from observational data
- Treated and non-treated groups
- Counterfactual – what would have happened to the treated, has they not received treatment
- Key assumption: that units selected into treatment and non-treatment groups have potential outcomes in both states

Identification Assumptions

Conditional Independence Assumption

"selection on observables" and participation is independent of outcomes once we control for observable characteristics (x)

Common Support Condition

we compare comparable individuals

PSA

- Treated Group

we have observed mean outcome under the condition of treatment $E(Y_1|W=1)$ and unobserved mean outcome under the condition of non-treatment $E(Y_0|W=1)$.

- Non-treated Group

we have both observed mean $E(Y_0|W=0)$ and unobserved mean $E(Y_1|W=0)$.

Under this framework, an evaluation of

$$E(Y_1|W=1) - E(Y_0|W=0)$$

can be thought as an effort that uses $E(Y_0|W=0)$ to estimate the counterfactual $E(Y_0|W=1)$. The central interest of the evaluation is not in $E(Y_0|W=0)$, but in $E(Y_0|W=1)$

Matching Estimators

All matching estimators are weighted estimators in which untreated units that closely resemble the treated ones (in terms of x) receive the highest weights.

$$\Delta_{ATT} = \frac{1}{n_i}\left[\sum_{i \in C} Y_i - \sum_{j \in C} w(\rho_i, \rho_j) Y_j\right]$$

where n_i are the number of treated individuals, and $w\left(\rho_i, \rho_j\right)$ is the weight placed on the *jth* untreated observation in constructing the counterfactual for the *ith* treated observation.

Different matching estimators differ in how they construct the weights *w(i,j)*.

Propensity Score

- Propensity score is the probability of taking treatment given a vector of observed variables.

$$\rho(x) = P[D = 1|X = x]$$

•Matching is a statistical approach that solves the evaluation problem by finding in a large group of nonparticipants those individuals who are similar to the participants in **all** relevant pre-treatment characteristics X.

•The simplest method of matching compares units with exactly the same values of x.

•Matching assumes that there is no selection bias based on unobserved characteristics

PSA Procedure

1. Get representative and comparable data on participants and nonparticipants
   (using the same survey and a similar time period)

2. Estimate the probability of program participation as a function of observable characteristics
   (using a logit or other discrete choice model)

3. Use predicted values from estimation to generate propensity score $\rho(x)$ for all treatment and comparison group members

4. Match Participants: Find a sample of non-participants with similar $\rho(x)$
   (Restrict samples to ensure common support)

5. Post-estimation

Double Bootstrap Procedure

1. Compute efficiency scores $\hat{\delta}_i = \hat{\delta}(x_i, y_i | \hat{\wp}) \; \forall \; i = \overline{1,n}$
2. Use maximum likelihood to obtain estimates $\hat{\beta}_i$ of β_i and an estimate $\hat{\sigma}_\varepsilon$ of σ_ε in the truncated regression of $\hat{\delta}_i$ on z_i (use $m < n$ if $\hat{\delta}_i > 1$).
3. Obtain n sets of bootstrap estimates of efficiency scores $\mathfrak{B}_i = \{\hat{\delta}_{ib}^*\}_{b=1}^{L_1}$, where L_1 is a number of bootstrap iterations.
4. For each $i = \overline{1,n}$ compute the bias corrected scores $\hat{\hat{\delta}}_i = \hat{\delta}_i - \widehat{BIAS}(\hat{\delta}_i)$ using the bootstrap estimates $\mathfrak{B}_i$ and the original estimates $\hat{\delta}_i$.
5. Use maximum likelihood to obtain estimates $\hat{\hat{\beta}}_i$ and $\hat{\hat{\sigma}}_\varepsilon$ from the truncated regression of $\hat{\hat{\delta}}_i$ on z_i.
6. Obtain a set of bootstrap estimates $\mathfrak{C} = \left\{\left(\hat{\hat{\beta}}_i^*, \hat{\hat{\sigma}}_\varepsilon^*\right)_b\right\}_{b=1}^{L_2}$, where L_2 is a number of bootstrap iterations.
7. Construct confidence intervals for each element in β_i and σ_ε using bootstrap values in $\mathfrak{C}$ and the original estimates of $\hat{\hat{\delta}}_i$ and $\hat{\hat{\sigma}}_\varepsilon$.

Simulation Setup

- Set $z_{i1} = 1$ and randomly choose $z_{ij} \sim N(\mu_z, \sigma_z^2)$ for $j = \overline{2, r}$.
- Generate a left truncated error term which will be used in the regression: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, where $\varepsilon_i = 1 - z_i\beta$.
- Set the inefficiency to be $\delta_i = z_i\beta + \varepsilon_i$.
- Inputs are: $x_{ij} \sim U(5,20)$ for $j = \overline{1, p}$.
- Total output: $y_i = \delta_i^{-1} \sum_{j=1}^{p} x_{ij}^k$
- Multiple outputs: y_i is split according to shares $\alpha_1 \sim U(0,1)$, $\alpha_j \sim U\left(0,1 - \sum_{j=1}^{q-2} \alpha_j\right)$ for $j = \overline{2, q-1}$.

  Therefore $y_{ij} = \alpha_j \delta_i^{-1} \sum_{j=1}^{p} x_{ij}^k$ for $j = \overline{1, q-1}$ and
$y_{iq} = \left(1 - \sum_{j=1}^{q-1} \alpha_j\right) \delta_i^{-1} \sum_{j=1}^{p} x_{ij}^k$.
- $\mu_z = 2, \sigma_z = 2, \sigma_\varepsilon = 1, k = {}^3/_4$ and $\beta_1 = \beta_2 = 0.5$

Truncated regression w/o Matching (One output, two inputs, and one external variable)						
n=100	True Parameter	Alg#2	90% Conf. Interval		95% Conf. Interval	
β_1	0.5	0.4689	-0.0449	1.2724	-0.1783	1.3857
β_2	0.5	0.4559	0.3189	0.5818	0.2840	0.5978
σ_ε	1	0.8281	0.4489	1.0991	0.3377	1.1547
n=400						
β_1	0.5	0.6278	0.3494	0.9587	0.3076	1.0834
β_2	0.5	0.4810	0.4052	0.5502	0.3975	0.5653
σ_ε	1	0.8722	0.6711	1.0374	0.5991	1.0592

Truncated regression w/ Matching						
n=100	True Parameter	Alg#2	90% Conf. Interval		95% Conf. Interval	
β_1	0.5	-0.4527	-1.7874	2.3940	-1.9135	3.3587
β_2	0.5	0.7567	0.2001	1.1241	-0.0810	1.1552
σ_ε	1	0.9419	-0.1902	1.5074	-0.5005	1.5403
n=400						
β_1	0.5	-0.0469	-0.9017	0.9453	-1.0891	2.1344
β_2	0.5	0.4849	0.2392	0.6721	0.0389	0.7156
σ_ε	1	1.2362	0.4834	1.7980	0.1066	1.8526

Truncated regression w/o Matching (Two outputs, three inputs, and three external variables)						
n=100	True Parameter	Alg#2	90% Conf. Interval		95% Conf. Interval	
β_1	0.5	-2.3162	-3.8200	-0.7897	-4.0074	0.1131
β_2	0.5	0.6291	0.4467	0.8238	0.4023	0.8509
β_3	0.5	0.6600	0.4170	0.8867	0.3344	0.9418
β_4	0.5	0.5959	0.3492	0.8019	0.2766	0.8588
σ_ε	1	2.2593	1.4939	3.0395	1.1348	3.1451
n=400						
β_1	0.5	-0.1017	-0.4710	0.2930	-0.5112	0.3955
β_2	0.5	0.5508	0.4869	0.6150	0.4793	0.6220
β_3	0.5	0.5442	0.4702	0.6134	0.4556	0.6273
β_4	0.5	0.5485	0.4863	0.6092	0.4689	0.6212
σ_ε	1	1.5576	1.3048	1.8129	1.2757	1.8402

Truncated regression w/ Matching						
n=100	True Parameter	Alg#2	90% Conf. Interval		95% Conf. Interval	
β_1	0.5	0.5843	-0.0020	1.2270	-0.0602	1.3579
β_2	0.5	0.3151	0.1874	0.4457	0.1484	0.4888
β_3	0.5	0.2845	0.1414	0.4377	0.1104	0.4442
β_4	0.5	0.3610	0.1851	0.5045	0.1554	0.5199
σ_ε	1	0.4383	0.2546	0.6477	0.2138	0.6698
n=400						
β_1	0.5	-0.5181	-1.1979	0.5533	-1.3821	0.9108
β_2	0.5	0.6100	0.3926	0.7551	0.3583	0.7749
β_3	0.5	0.5768	0.3619	0.7196	0.3382	0.7595
β_4	0.5	0.5985	0.4209	0.7624	0.3607	0.7751
σ_ε	1	1.6368	1.0467	2.0889	0.9284	2.1969

Texas Electricity Market

- Electricity Generators in ERCOT
- 1999-2002 (69 power plants)

- Net Generation, Exclusive of Plant Use (KWh)
- Total Installed Cap (MW)
- Average Number of Employees (Employees)
- Fuel ($)
- Plant Hours Connected to Load (Hours)
- Total Cost ($)
- Total Production Expenses ($)

DEA

- CRS
- Total Cost and Total Production Expenses

  in $10000
- All variables in logs

Regression

• Fuel	0.8895
• Connected to Load	-3.9230
• Total Cost	0.6233
• Total Production Expenses	0.7126
• Sigma	0.8084

Further Research

- Multiple treatments
  - One Variable
  - Different Variables

Conclusions

- PSA takes into the account all the covariates in before the first stage
- PSA provides good supplementary information to the conventional frontier methods
- May be used together with DEA/SFA