Introducing Environmental Variables in Two-Stage DEA and Conditional Efficiency Model with Test of the Separability Assumption

HackMd Link:
https://hackmd.io/o3P9mzp8SleroDl1uD9fhQ?view
(https://hackmd.io/o3P9mzp8SleroDl1uD9fhQ?view)

1. Introduction

1.1. Motivation

Data envelopment analysis (DEA) esitimates efficiency of each decision making unit (DMU); however, some exogenous factors might affect firms' performance. This project try to account those factors, i.e., environmental variables, in DEA models (Two-Stage DEA and Conditional Efficiency Model) under different model assumptions (separability).

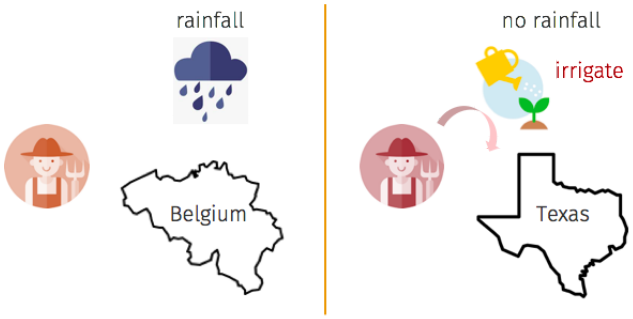
1.2. Background

Here, we discuss two main topics about this project, which are "environmental variables" and "separability".

- **Environmental Variables**

In any production unit, factors of the external environment generally influence the ability of management to transform inputs into outputs. These factors are the environment variables.

For example, in agricultural applications, one might use rainfall as an environmental variable—farmers in Belgium do not irrigate their crops, but farmers in west Texas must do so.



- **Separability**

SW note (pp. 35–36) that their Assumptions A1–A2 imply a "separability" condition. Specifically, by "separability", we mean that the support of the output variables does not depend on the environmental variables in Z.

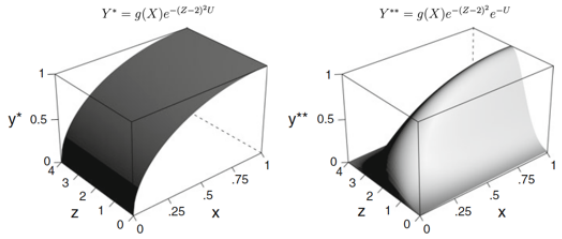
To illustrate this condition, consider the two DGPs given by

$$Y^* = g(X)e^{-(Z-2)^2U}$$

and

$$Y^* = g(X)e^{-(Z-2)^2}e^{-U}$$

where
$g(X) = (1-(X-1)^2)^{1/2}, X \in [0,1], Z \in [0,4], U \geq 0$
(one-sided inefficiency).



The left figure shows the production frontier satisfies the separability assumption, while the right figure shows the production frontier does not satisfy the separability assumption since the production frontier changes while shifting the environmental variable Z.

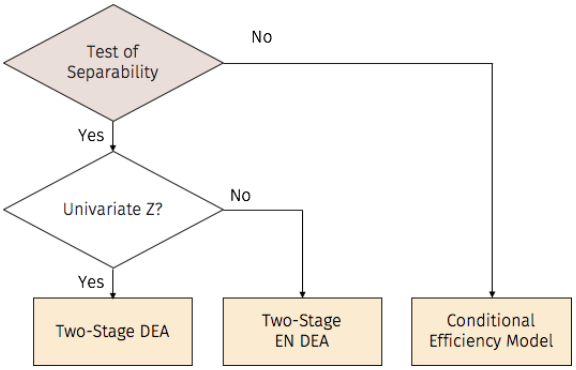
1.3. Problem Definition

This project aim to model the environmental variables and evaluate the efficiencies. (consider whether the " separability" assumption is satisfied or not)

2. Proposed Framework

- We proposed a framework to solve above problem which includes
  - Test of the Separability Assumption
  - Two-stage DEA
  - Conditional Efficiency Model

  We first test the separability assumption of the given data. If the separability assumption is not violated, then we construct a two-stage DEA model (construct a two-stage elastic-net DEA model if environmental variables are multivariate). If the separability assumption isn't satisfied, then we construct a conditional efficiency model.



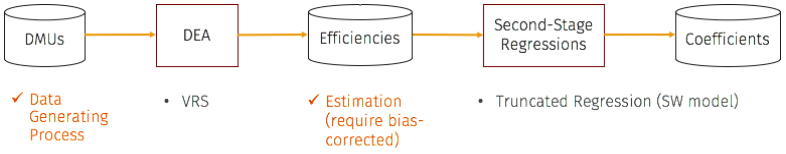
In order to analyze by above framework, the following section will go through these two different models.

3. Two-Stage DEA

In this section, a two-stage estimation procedures is introduced. The technical efficiency is estimated by data envelopment analysis (DEA) or free disposal hull (FDH) estimators in the first stage, and the resulting efficiency estimates are regressed on some environmental variables in a second stage.

3.1. Introduction

The figure below shows how a general two-stage DEA is constructed. Specifically, in the first stage, efficiencies can be estimated by a output-oriented VRS model; in the second stage, a boostrap method is applied to construct a bias-corrected estimates of efficiencies with a truncated regression regressing on environmental variables (Simar-Wilson model). Therefore, we can get the efficiencies and the coefficients of given environmental variables. We will demonstrate this procedures via a Monte-Carlo experiment following a specific data generating process (DGP) in the later subsection.



3.2. Simar-Wilson (SW) Model

- Model Assumption
  - Generally, in the two-stage DEA, the second stage truncated regression model is assume as following formulation,

    $$\delta_i = \Psi(\boldsymbol{Z}_i, \boldsymbol{\beta}) + \varepsilon_i \geq 1$$

    , where
    δ_i : output efficiency measure ($\delta_i \geq 1$ by definition) (Farrell, 1957)
    $\boldsymbol{Z}_i$: environmental variables
    $\boldsymbol{\beta}$: parameters (coefficients)
    Ψ : function of environmental variables and parameters
    ε_i : the part of inefficiency not explained by $\boldsymbol{Z}_i$, it's assumed to be distributed $N(0, \sigma_\varepsilon^2)$ with left-truncation at $1 - \Psi(\boldsymbol{Z}_i, \boldsymbol{\beta})$
  - Unfortunately, the δ_i are not observed. Therefore, DEA estimates $\hat{\delta}_i$ from the first stage estimation are used to replace the unobserved δ_i, which is

    $$\hat{\delta}_i = \Psi(\boldsymbol{Z}_i, \boldsymbol{\beta}) + \varepsilon_i \geq 1.$$

Then, we show how the to make the inference of above model.

- Inference
  - As SW note, inference is problematic due to the fact that $\hat{\delta}_i$ has replaced the unobserved δ_i, and while the $\hat{\delta}_i$ consistently estimate the δ_i, the DEA estimators converge slowly, at rate $n^{-2/(p+q+1)}$, and are biased. Consequently, the inverse of the negative Hessian of the log-likelihood does not consistently estimate the variance of the ML estimator of $\boldsymbol{\beta}$.
  - SW show how bootstrap methods can be used to construct bias-corrected estimates $\hat{\delta}_i$ of the unobserved δ_i and make inference about $\boldsymbol{\beta}$ (with confidence in).
  - We will introduce the algorithm of such bootstrap procedure in the next subsection.

- Contribution
  - The algorithms proposed by SW contribute mainly in two parts.
    - SW define a statistical model where truncated regression yields consistent estimation of model features.
    - SW demonstrated that conventional, likelihood-based approaches to inference are invalid and developed a bootstrap approach that yields valid inference in the second-stage regression.

3.3. Monte Carlo Experiment

Here, we conducted a Monte Carlo experiment to examine the performance of the algorithm to inference in the secondstage regression. The data is generated from a known process and applied our bootstrap algorithm on each of M Monte Carlo trials.

- Data Generating Process
  - The data was generated as following procedure:
    (1) Draw $z_{ij} \sim N(\mu_z, \sigma_z^2)$ for $j = 1, \ldots, r$
    (2) Draw $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ with left-truncated at $1 - \boldsymbol{Z}_i\boldsymbol{\beta}$
    (3) Set $\delta_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \varepsilon_i$
    (4) Draw $x_{ij} \sim U(6, 16)$
    (5) Set $y_i = \delta_i^{-1} \sum_{j=1}^{p} x_{ij}^{3/4}$

- The parameter and Monte-carlo experiment settings is as follow,
  - $Z : \mu_Z = 2, \sigma_Z^2 = 2, r = 1$
  - $\varepsilon : \sigma_\varepsilon^2 = 1$
  - $\beta : \beta_0 = 0.5, \beta_1 = 0.5$
  - $N = 100, 400$ (sample size)
  - $M = 100$ (# of Monte-Carlo trials)
  - $L = 100$ (# of bootstraps)
  - $\alpha = 0.2, 0.1, 0.05, 0.01$ (significance levels) we compare two different sample sizes of data with different significance levels.

- Algorithm
  - The pseudocode of two-stage DEA alogrithm is as follow,

[1] Using the original data in $\mathscr{S}_n$, compute $\widehat{\delta}_i = \widehat{\delta}(x_i, y_i | \widehat{\mathscr{P}}) \forall i = 1, \ldots, n$ using (10).
[2] Use the method of maximum likelihood to obtain an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ as well as an estimate $\widehat{\sigma}_\varepsilon$ of σ_ε in the truncated regression of $\widehat{\delta}_i$ on $\boldsymbol{z}_i$ in (13) using the $m < n$ observations where $\widehat{\delta}_i > 1$.
[3] Loop over the next three steps ([3.1]–[3.3]) L times to obtain a set of bootstrap estimates $\mathscr{A} = \{(\widehat{\boldsymbol{\beta}}^*, \widehat{\sigma}_\varepsilon^*)_b\}_{b=1}^L$:
[3.1] For each $i = 1, \ldots, m$, draw ε_i from the $N(0, \widehat{\sigma}_\varepsilon^2)$distribution with left-truncation at $(1 - \boldsymbol{z}_i \widehat{\boldsymbol{\beta}})$.[7]
[3.2] Again for each $i = 1, \ldots, m$, compute $\delta_i^* = \boldsymbol{z}_i \boldsymbol{\beta} + \varepsilon_i$.
[3.3] Use the maximum likelihood method to estimate the truncated regression of δ_i^* on $\boldsymbol{z}_i$, yielding estimates $(\widehat{\boldsymbol{\beta}}^*, \widehat{\sigma}_\varepsilon^*)$.
[4] Use the bootstrap values in $\mathscr{A}$ and the original estimates $\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_\varepsilon$ to construct estimated confidence intervals for each element of $\boldsymbol{\beta}$ and for σ_ε as described below.

- Results: estimated coverage of confidence intervals
  - We ran 100 Monte Carlo trials and compute the proportion among the 100 Monte Carlo experiments where the estimated confidence interval covers the true value of β_0, β_1, and σ_ε^2 at different significance levels. The table below shows the proportion of esimated coverge of confidence intervals which provide valid inference.
  - One can easily use the package `main.py` to conduct the Monte-Carlo experiment.

n	Param.	Significance level			
		0.20	0.10	0.05	0.01
100	β_0	0.61	0.73	0.78	0.85
	β_1	0.61	0.70	0.74	0.87
	σ_ε	0.48	0.66	0.77	0.83
400	β_0	0.55	0.72	0.79	0.89
	β_1	0.61	0.74	0.81	0.89
	σ_ε	0.50	0.69	0.76	0.85

4. Conditional Efficiency Model

In this section, a nonparametric frontier model under probabilistic representation is introduced. We also consider the environmental factors which may affect neither input or output side but the whole production

prcess. Additionally, the order-m efficiency measure is developed via bootstrapping method. Finally, python implementation and a toy example are proposed.

4.1. Formulation of the Production Process

Denote a set of inputs $x \in \mathbb{R}_+^p$, a set of outputs $y \in \Psi \in \mathbb{R}_+^q$ and $\Psi = \{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\}$, the possibile produciton set. The frontier (i.e. boundaries of Ψ) becomes the measurment of the efficiency score. The Farrell of input-oriented efficiency score for a DMU at the level (x, y) is defined as:

$$\lambda(x, y) = \inf\{\lambda \mid (\lambda x, y) \in \Psi\}.$$

Note that the production process can be reformulated in a probabilistic way. Denote $H_{XY}(\cdot, \cdot)$ the probability for a unit operating at the level (x, y) to be dominated, where

$$\begin{aligned} H_{XY}(x, y) &= \text{Prob}(X \leq x, Y \geq y) \\ &= \text{Prob}(X \leq x \mid Y \geq y)\text{Prob}(Y \geq y) \\ &= F_{X|Y}(x|y)S_Y(y), \end{aligned}$$

Therefore, the input oriented efficiency score $\lambda(x, y)$ is defined for all y with $S_Y(y) > 0$ as

$$\lambda(x, y) = \inf\{\lambda \mid F_{X|Y}(\lambda x \mid y) > 0\} = \inf\{\lambda \mid H_{XY}(\lambda x, y) > 0\}.$$

However, since $F_{X|Y}$ is unknown and can't be obtained, we may use the empirical version $\hat{F}_{X|Y,n}$ to replace this term:

$$\hat{F}_{X|Y,n}(x \mid y) = \frac{\sum_{i=1}^n I(X_i \leq x, Y_i \geq y)}{\sum_{i=1}^n I(Y_i \geq y)},$$

where $I(\cdot)$ is the indicator function.

4.2 Efficiency Estimator

Free Disposal Hull (FDH) and Variable Return to Scale (VRS) models are particularly used to estimate efficiencies. We may consider different production sets when using these estimators. $\hat{\Psi}_{FDH}$ and $\hat{\Psi}_{VRS}$ are represented as:

$$\hat{\Psi}_{FDH} = \{(x, y) \in \mathbb{R}_+^{p+q} \mid x \geq x_i, y \leq y_i, i = 1, \ldots, n\}.$$

As for VRS estimator, $\hat{\Psi}_{VRS}$ is obtained by the convex hull of $\hat{\Psi}_{FDH}$:

$$\hat{\Psi}_{VRS} = \{(x, y) \in \mathbb{R}_+^{p+q} \mid y \leq \sum_{i=1}^n \gamma_i y_i; x \geq \sum_{i=1}^n \gamma_i x_i; \sum_{i=1}^n \gamma_i = 1; \gamma_i \geq 0, i = 1, \ldots, n\}.$$

Therefore, we may have the efficiency estismators $\hat{\lambda}_{FDH}(x,y) = \inf\{\lambda \mid (\lambda x, y) \in \hat{\Psi}_{FDH}\}$ and $\hat{\lambda}_{VRS}(x,y) = \inf\{\lambda \mid (\lambda x, y) \in \hat{\Psi}_{VRS}\}$ under FDH and VRS scenarios, respectively.

4.3 Conditional Measures of Efficiency

If we want to take environmental factors into consideration, joint distribution of (X, Y) can be revised to add a new condition. Denote $Z \in \mathbb{R}^r$ the environmental facotrs. Ψ^z can be represented as $H_{X,Y|Z}(x,y \mid z) = \text{Prob}(X \leq x, Y \geq y \mid Z = z) = F_{X|Y,Z}(x \mid y, z) S_{Y|Z}(y \mid z)$. Thus, corresponding conditional efficiency $\theta(x, y \mid z)$ is defined as

$$\lambda(x, y \mid z) = \inf\{\theta \mid F_{X|Y,Z}(\lambda x \mid y, z) > 0\}$$

Similarly, we face the unknown distribution $F_{X|Y,Z}$ again, so the empirical distribution with kernel density estimator for the environmental variables is considered:

$$\hat{F}_{X|Y,Z,n}(x \mid y, z) = \frac{\sum_{i=1}^{n} I(X_i \leq x, Y_i \geq y) K((z - z_i)/h)}{\sum_{i=1}^{n} I(Y_i \geq y) K((z - z_i)/h)},$$

where $K(\cdot)$ is the kernel function and h is the bandwidth of appropriate size. It indicats that the kernel should be with compact support (i.e., $K(u) = 0$ if $|u| > 1$, as for the uniform, triangle, epanechnikov or quartic kernels). The issue of the chosen of the bandwidth h is also discussed in Daraio, C., & Simar, L. (2005). In general, one can use cross-validation to adjust h.
The conditional FDH and VRS efficiency score can thus be defined as:

$$\begin{aligned}
\hat{\lambda}_{FDH}(x, y \mid z) &= \inf\{\lambda \mid (\lambda x, y) \in \hat{\Psi}^z_{FDH}\} \\
&= \inf\{\lambda \mid \hat{F}_{X|Y,Z,n}(\lambda x \mid y, z) > 0\} \\
&= \min_{\{i \mid Y_i >= y, |Z_i - z| \leq h\}} \{\max_{j=1,\ldots,p} \frac{X_i^j}{x^j}\}
\end{aligned}$$

$$\begin{aligned}
\hat{\lambda}_{VRS}(x, y \mid z) &= \inf\{\lambda \mid (\lambda x, y) \in \hat{\Psi}^z_{VRS}\} \\
&= \inf\{\lambda \mid y \leq \sum_{\{i \mid z-h \leq z_i \leq z+h\}} \gamma_i y_i; \lambda x \geq \sum_{\{i \mid z-h \leq z_i \leq z+h\}} \gamma_i x_i; \sum_{\{i \mid z-h \leq z_i \leq z+h\}} \gamma_i = 1; \gamma_i \geq 0\}
\end{aligned}$$

4.4 Order-m Frontiers and Efficiency Scores

Since both FDH and VRS efficiency estimators are sensitive to extreme DMUs and outliers, one can get more robust efficiency estimations via order-m frontiers. Formally, for a given level of output y, we consider m i.i.d. random samples $X_1, \ldots, X_m$ generated by the conditional p-

variate function $F_{X|Y}(\cdot \mid y)$ and obtain the random production set of order-m for units producing more than y, defined as:

$$\tilde{\Psi}_m(y) = \{(x, y') \in \mathbb{R}_+^{p+q} \mid x \geq X_i, y' \geq y, i = 1, \ldots, m\}.$$

Thus, the corresponding order-m efficiency score is defined as:

$$\lambda_m(x, y) = E_{X|Y}(\tilde{\lambda}_m(x, y) \mid Y \geq y),$$

where $\tilde{\lambda}_m(x, y) = \inf\{\lambda \mid (\lambda x, y) \in \tilde{\Psi}_m(y)\}$ and the efficiency estimator $\hat{\lambda}$ can be either FDH or VRS.
Note that the expectation term is complicated to compute; thus, the practical computations algorithm are proposed in the next section.

4.5 Practical Computation Algorithm

This section is quoted from Daraio, C., & Simar, L. (2007). Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. *Journal of productivity analysis*, 28(1), 13-32.

a. Order-m FDH Efficiency

1. For a given y, draw a sample of size m with replacement among those X_i such that X_i s.t. $Y_i \geq y$ and denote this sample by $(X_{1,b}, \ldots, X_{m,b})$
2. $\tilde{\lambda}_m^b(x, y) = \min_{i=1,\ldots,m}\{\max_{j=1,\ldots,p} \frac{X_{i,b}^j}{x^j}\}$
3. Redo 1. - 2. for $b = 1, \ldots, B$, where B is large.
4. Finally, $\hat{\lambda}_m(x, y) \approx \frac{1}{B} \sum_{b=1}^{B} \tilde{\lambda}_m^b(x, y)$

b. Order-m VRS Efficiency

1. For a given y, draw a sample of size m with replacement among those X_i such that X_i s.t. $Y_i \geq y$ and denote this sample by $(X_{1,b}, \ldots, X_{m,b})$
2. Solve the following linear program:
   $\tilde{\lambda}_m^b(x, y) = \inf\{\lambda \mid \lambda x \geq \sum_{i=1}^{m} \gamma_i X_{i,b}; \sum_{i=1}^{m} \gamma_i = 1; \gamma_i \geq 0, i = 1, \ldots, m\}$
3. Redo 1. - 2. for $b = 1, \ldots, B$, where B is large.
4. Finally, $\hat{\lambda}_m(x, y) \approx \frac{1}{B} \sum_{b=1}^{B} \tilde{\lambda}_m^b(x, y)$

c. Order-m conditional FDH Efficiency

1. For a given y, draw a sample of size m with replacement and with a probability $\frac{K((z-z_i)/h)}{\sum_{j=1}^{n} K((z-z_j)/h)}$ among those X_i such that X_i s.t. $Y_i \geq y$ and denote this sample by $(X_{1,b}, \ldots, X_{m,b})$

2. $\tilde{\lambda}_m^b(x,y) = \min_{i=1,\ldots,m}\{\max_{j=1,\ldots,p} \frac{X_{i,b}^j}{x^j}\}$
3. Redo 1. - 2. for $b = 1, \ldots, B$, where B is large.
4. Finally, $\hat{\lambda}_m(x,y) \approx \frac{1}{B}\sum_{b=1}^B \tilde{\lambda}_m^b(x,y)$

d. Order-m conditional VRS Efficiency

1. For a given y, draw a sample of size m with replacement and with a probability $\frac{K((z-z_i)/h)}{\sum_{j=1}^n K((z-z_j)/h)}$ among those X_i such that X_i s.t. $Y_i \geq y$ and denote this sample by $(X_{1,b}, \ldots, X_{m,b})$
2. Solve the following linear program:
   $\tilde{\lambda}_m^b(x,y) = \inf\{\lambda \mid \lambda x \geq \sum_{i=1}^m \gamma_i X_{i,b}; \sum_{i=1}^m \gamma_i = 1; \gamma_i \geq 0, i = 1, \ldots, m\}$
3. Redo 1. - 2. for $b = 1, \ldots, B$, where B is large.
4. Finally, $\hat{\lambda}_m(x,y) \approx \frac{1}{B}\sum_{b=1}^B \tilde{\lambda}_m^b(x,y)$

4.6 Python Implementation and Numerical Study

In this project, we implement differents types of efficiency estimation methods including FDH and VRS model under different scenarios as below:

Frontier	Unconditional	Conditional
Full	Without considering Z Use all datapoints	Consider Z Use all datapoints
Partial (order-m)	Without considering Z Use bootstrapping	Consider Z Use bootstrapping

One can easily use the package `EfficiencyCalculator.py` to get the efficiency estimations.

```
cal = EfficiencyCalculator(x, y)
cal.set_environmental_variables(z)
cal.set_bandwidth(h=0.01)
cal.set_kernel(kernel='triangular')
cal.get_full_efficiency(dmu=6, conditional=True, method='VRS')
cal.get_partial_efficiency(dmu=6, conditional=True, method='VRS')
```

Note that

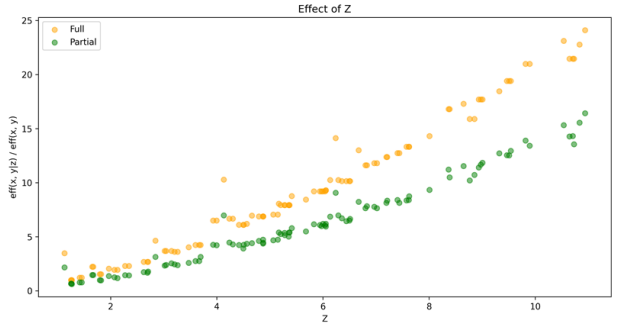
- Currently, it only supports triangular kernel.
- `dmu` parameter indicates a certain unit index in the x, y, z numpy arrays.
- `method` could be `VRS` or `FDH`.
- `conditional` colud be `True` or `False`.

In the following cases, we assume all units with the same output units; thus, the DMU with less input units will be regarded as more efficient.
The data generating process for two cases and some parameters are shown below:

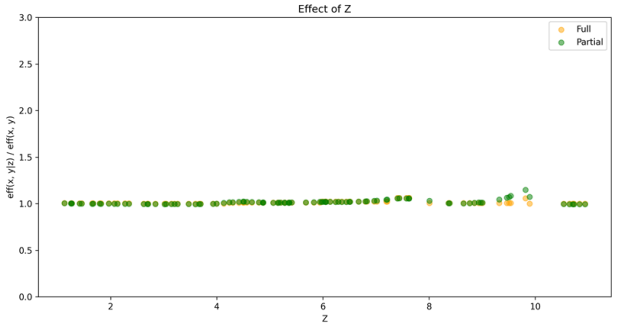
- Case I: $X = Z^{3/2} + \epsilon$
- Case II: $X = 5^{3/2} + \epsilon$,
- $n = 100$
- $m = 25$
- $B = 200$

where ϵ is a noise term.
We calculate the full frontier and partial frontier (order-m) FDH efficiencies for these 100 units and perform the scatter plots against the value of univariate environmental varialbe z. The y-axis indicates the conditional efficiency divided by unconditional efficiency, i.e., $\lambda(x, y \mid z)/\lambda(x, y)$. Therefore, the value can illustrates the effect from the undesired environmental factor.

Case I:
The trends for full and partial efficiency estimation are both increasing, which indicates that z is unfavorable.



Case II:
It shows that z is independent of the efficiency estimation, which meets our expectation considering the corresponding data generating process.



The codes can be found in `efficiency_example.ipynb` .

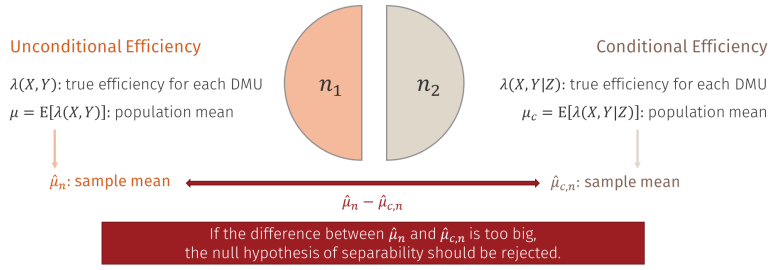
5. Test of the Separability Assumption

In our proposed framework from section 2, the test of separability assumption is extremely crucial. Afterward, one can decide which model is suitable to estimate the efficiency and the effect from the environmental variables. However, the methodology is complicated in the original paper. Thus, we only introducte the basic idea in this project. Details can be investigated in Daraio, C., Simar, L., & Wilson, P. W. (2018). Central limit theorems for conditional efficiency measures and tests of the 'separability'condition in non-parametric, two-stage models of production. *The Econometrics Journal*, 21(2), 170-191.

We first state the null and the alternative hypothesis:

$$
\begin{aligned}
H_0 &: \text{Separability is hold.} \\
H_1 &: \text{Separability is violated.}
\end{aligned}
$$

The main idea for the test is that we can divide the samples into two groups. One is used to estimate unconditional efficiency and the other is used to estimate conditional one. Under the null, the two population mean should be similar. Therefore, we may use sample mean to construct a test statistic.



The table below shows the efficiency estimation under different scenarios. The first row is the unconditional efficiency, and the last row is the conditional one. Note that the esitmator $\hat{\lambda}(x, y \mid z)$ targets $\Psi^{z,h}$ instead of Ψ^z. However, as $h \to 0$, these two terem will converge to the same value. In addition, B is the bias term and R is the remainder term, which will vanish under some conditions. The details please refer to the original paper in section 5.

Production Set	True Efficiency	Estimated Efficiency	Sample Mean	Population Mean
$\Psi = \{(X,Y)\}$	$\lambda(x,y)$	$\hat{\lambda}(x,y)$	$\hat{\mu}_n$	μ
			$\sqrt{n}(\hat{\mu}_n - B - R) \xrightarrow{\mathcal{L}} N(\mu, \sigma^2)$	
$\Psi^z = \{(X,Y)\|Z = z\}$	$\lambda(x,y\|z)$			
$\Psi^{z,h} = \{(X,Y)\|\|Z - z\| \le h\}$	$\lambda^h(x,y\|z)$	$\hat{\lambda}(x,y\|z)$	$\hat{\mu}_{c,n_h}$	μ_c^h
			$\sqrt{n_h}(\hat{\mu}_{c,n_h} - B^* - R^*) \xrightarrow{\mathcal{L}} N(\mu_c^h, \sigma_c^{2,h})$	

Finally, the test statistic T can be established from the two subsample groups as:

$$
T_{1,n} = \frac{(\hat{\mu}_{n_1} - \hat{\mu}_{c,n_{2,h}}) - (\hat{B}_{\kappa,n_1} - \hat{B}_{\kappa,n_{2,h}})}{\sqrt{(\hat{\sigma}^2_{n1}/n_1) + (\hat{\sigma}^{2,h}_{c,n_2}/n_{2,h})}} \xrightarrow{\mathcal{L}} N(0,1)
$$

Given a significance level α, one can reject H_0, if $1 - \phi(T_{1,n}) < \alpha$, where $\phi(\cdot)$ is the standard normal cdf. To sum up, the test procedure is established from the central limit theorem for conditional and unconditional efficiency measures, and the corresponding test statistic is constructed by the two groups separated from the whole samples. Once the null hypothesis is rejected, one can't use two-stage DEA model to evaluate the effect from the environmental variables. Thus, conditional efficiency model is more suitable under this scenario.

6. Conclusion

6.1. Contribution

In this project, our proposed framework is able to

- Test the separability condition
- Estimate efficiencies considering environmental variables Z
- Deal with multivariate environmental variables Z

6.2. Limitation

In this project, our proposed framework is unable to

- Estimate the influence of environmental variables Z if the separability does not satisfy
- Estimate efficiencies and coefficients of environmental variables Z in two-stage DEA simultaneously

7. Future work

Here, we sugguest three different aspects with the corresponding issues that can be study in further research.
(1) Environmental variables

- Nonlinear or nonparametric truncated regression in the second stage
- Dependency within the two stages (Two-Stage DEA) $\Longrightarrow$ Does the information from the first stage totally used in the second stage?

(2) Number of variables

- Feature selection simultaneously (inputs, outputs, environmental variables)

(3) CLT for conditional efficiency measures

- Hypothesis testing for small sample size

References

Cazals, C., Florens, J. P., & Simar, L. (2002). Nonparametric frontier estimation: a robust approach. *Journal of econometrics*, 106(1), 1-25.
Daraio, C., & Simar, L. (2005). Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of productivity analysis*, 24(1), 93-121.
Daraio, C., & Simar, L. (2007). Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. *Journal of productivity analysis*, 28(1), 13-32.
Bădin, L., Daraio, C., & Simar, L. (2012). How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research*, 223(3), 818-833.

Daraio, C., Simar, L., & Wilson, P. W. (2018). Central limit theorems for conditional efficiency measures and tests of the ‘separability’condition in non-parametric, two-stage models of production. *The Econometrics Journal*, 21(2), 170-191.