SSE-DP-2022-4

A Statistical Data Envelopment Analysis

Naoto Kunitomo
(The Institute of Statistical Mathematics)
and
Yu Zhao
(Tokyo University of Science)

November 2022
(May 2023, Revised)

A Statistical Data Envelopment Analysis *

Naoto Kunitomo †
and
Yu Zhao ‡

November 19, 2022
May 15, 2023 (Revised)

Abstract

In operations research and management sciences, the data envelopment analysis (DEA) has been known as one of important tools. We develop a statistical data envelopment analysis (SDEA), which seems to be new to operations research as well as statistical literature. We first consider the basic statistical DEA model that the observed data is the sum of an increasing concave function of inputs and a non-positive random noise. The noise term can be interpreted as the inefficiency of inputs-output relationships. The purpose of data analysis is to estimate the unknown function, called the efficiency frontier, nonparametrically based on the set of observed data of inputs and outputs. The key idea is to use the statistical methods of regression analysis and the statistical extreme value theory (SEVT). We report an empirical analysis on the life-insurance industry in Japan as an application.

Key words

Statistical Data Envelopment Analysis, Inefficiency, Regression Envelopment, Type-II Extreme-Value Distribution, Life-insurance Industry in Japan

*Version 2023-5-15. This is a revision of the discussion paper of SSE-DP-2022-4.pdf at https://stat-expert.ism.ac.jp/wp/wp-content/uploads/2023/02/ SSE-DP-2022-4.pdf, the Institute of Statistical Mathematics (ISM), Tokyo, Japan. This research has been supported by a project of Consortium for training experts in statistical sciences at ISM and it has been also supported by JSPS-Grant 22K01428.

†The Institute of Statistical Mathematics, JAPAN.

‡Tokyo University of Science, JAPAN.

1 Introduction

In operations research and management sciences, the data envelopment analysis (DEA) has been known as one of important tools. See Cooper, Seiford, and Tone (2007) for the details of existing known methods and history in operations research. The DEA in operations research is ususally based on the mathematical programming techniques. In economics, on the other hand, the parametric statistical estimation method of production frontier has been known since Aigner, Lovell, and Schmidt (1977). It is also related to the cost function and the problem is fundamental in micro-econometrics. They proposed a parametric nonlinear regression model with truncated distributions such as half-normal for noise term. By using the maximum likelihood estimation (MLE) method, they measured the parametric production function, which is an important tool in micro-econometrics. Some detail of econometric studies on the estimation problem of stochastic frontier functions has been explained in Chapter 17 of Green (2003). The main purposes of these two methods in operation research and econometrics are similar, but their traditional approaches and mathematical techniques to solve the similar problems are quite different.

In this paper, we develop a statistical data envelopment analysis (SDEA), which seems to be new to operations research literature as well as econometrics and statistical sciences. We first consider the basic statistical DEA model that the observed data is the sum of an increasing concave function and a pon-positive random noise. The random noise term can be interpreted as the inefficiency of inputs-output relationships. The purpose of statistical data analysis is to estimate the unknown function, called the efficiency frontier, nonparametrically based on the set of observed data. The key idea of the present work is to use the statistical methods of the regression analysis, and the statistical extreme value theory (SEVT) to estimate the unknown envelop function. When the sample size is not large, we have found that the estimation method based on the SEVT method may not be satisfactory in some cases. As the first estimation method, we shall use an estimation method based on the linear regression, which is quite simple and straightforward. However, we find that it has some possible efficiency loss in estimation when the sample size is large and it can be improved. Then, we shall introduce the second estimation method based on the SEVT method. We shall show that the order of second estimation method of unknown parameters is faster than the first estimation method in some situations. We also discuss the case when we have measurement errors as well as inefficiencies in noisy observed data sets.

The main purpose of this paper is to develop a new statistical approach to the DEA problem and some theoretical results. We also report an empirical analysis of the life insurance industry in Japan as an application. Since the number of data is about 40, which is quite small as the DEA problem, we have applied the regression-based method to this application. Since our approach is not along the traditional

approach in operations science and management sciences, first we explain the basic case, and then we generalize the simple formulation to more general cases.

The reminder of this paper is organized as follows. In Section 2, we discuss the formulation of SDEA and introduce the first estimation method in the simple case. In Section 3, we introduce the second estimation method for the case when the sample size is large. In Section 4, we discuss the relation of our SDEA model , the type-II extreme value distribution and the SEVT method. In Section 5, we generalize the basic SDEA method when we have several explanatory variables. In Section 6, we discuss the problem of measurement errors in the analysis of efficient frontier. In Section 7, we report an empirical study of the SDEA method for the life-insurance industry in Japan. Finally, in Section 8, we provide some concluding remarks.

2 A New Approach of SDEA

2.1 Statistical Data Envelopment Analysis

We formulate our problem as the non-parametric estimation of a statistical DEA model. Let the output level and input-level be Y and X, respectively, which take non-negative values. We assume that the efficient frontier function $h(\cdot)$ is smooth and twice-differentiable with $h^{'} > 0$ and $h^{''} < 0$, and the input variable X is fixed in this paper. (We usually consider the case when we only know that $f(\cdot)$ is a concave function in applications.) Let also the random variable U representing the inefficiency term from the (unknown) efficient frontier function, and we assume the relation

$$Y = h(X) + U \ \ (U \leq 0) \ . \tag{2.1}$$

In the standard DEA, both X and Y take any real numbers, and in real applications we only observed a finite number of data on X and Y. We use N as the sample size.

Let Y_i $(i = 1, \cdots, N)$, X_i $(i = 1, \cdots, N)$ are the observed output and input levels, which are non-negative, and $h_m(X)$ is an increasing concave piece-wise linear frontier function of the input level X as

$$h_m(x) = a_k + b_k x \ \ (x \in \mathrm{I}_k^{(m)} \, ; k = 1, \cdots, m) \ , \tag{2.2}$$

where $\mathrm{I}_k^{(m)} = (w_1^{(k)}, w_2^{(k)}]$ ($w_1^{(k)} \leq w_2^{(k)}$), $0 \leq w_1^{(1)} < \cdots < w_1^{(m)}$ and $0 \leq w_2^{(1)} < \cdots < w_2^{(m)}$.

In this study, we restrict our formulation to the case when X_i is a bounded deterministic variablIe and $w_1^{(1)} \leq X_1 \leq X_2 \leq \cdots \leq X_N \leq w_2^{(m)}$. Because of concavity, we impose the monotonicity restrictions on coefficients such that

$$0 \leq a_1 \leq \cdots \leq a_m \ , \ b_1 \geq \cdots \geq b_m \geq 0 \ . \tag{2.3}$$

Let also U_i $(i = 1, \cdots, N)$ is a sequence of i.i.d. random variables, which take non-positive values. As a typical case, U_i follows the negative exponential distribution such that for some positive $\lambda > 0$,

$$F(u) = P(U_i \leq u) = \exp[\lambda u] \ (u \leq 0) \ . \tag{2.4}$$

The basic statistical model is given by

$$Y_i = h_m(X_i) + U_i \ \ (i = 1, \cdots, N) \ . \tag{2.5}$$

The important feature of this representation is the restrictions that $h_m(X_i)$ is in the class of non-decreasing piece-wise linear concave function and U_i takes only non-positive real values. The efficient frontier function $h(X)$, which is the main interest of investigation, but it is unknown for researchers. This problem has been well known as the DEA model in operations research and there have been numerous applications. Also in econometrics, there has been some literature such as the econometric estimation of production frontier. (See Green (2003), for instance.)

Given a finite number of data sets (X_i, Y_i) $(i = 1, \cdots, N)$, it is only possible to estimate the unknown function $h_m(x)$ when $m = m_N$ is less than N. We divide the intervals $\mathrm{I}_k^{(m)}$ $(k = 1, \cdots, m)$ such that $\bigcup_{k=1}^m \mathrm{I}_k^{(m)} = (w_1^{(1)}, w_2^{(m)}]$ and we denote n_k as the number of data in $\mathrm{I}_k^{(m)} = (w_1^{(k)}, w_2^{(k)}]$ with

$$\sum_{k=1}^{m} n_k \geq N \ . \tag{2.6}$$

We allow the case when $\sum_{k=1}^m n_k > N$, which means data sets are overlapped in intervals.

In the present study, we shall consider the case when the input variable X is fixed (or there are several fixed input variables), and the bounds of intervals are known in advance. However, for example, it may be possible to pick intervals for X randomly. We conjecture that some efficient estimation methods for finite data would be developed, which is beyond the scope of the present work.

When both n_k $(k = 1, \cdots, m)$ and m are large, it is possible to develop the asymptotic theory for the estimation methods. In the following, wel investigate the (consistent) statistical estimation methods of the piece-wise linear function $\hat{h}_m$ such that as $m \to +\infty$ (and $n_k \to +\infty$).

$$\sup_x |\hat{h}_m(x) - h(x)| \xrightarrow{p} 0 \ . \tag{2.7}$$

It is because

$$\sup_x |\hat{h}_m(x) - h(x)| \leq \sup_x |\hat{h}_m(x) - h_m(x)| + \sup_x |h_m(x) - h(x)| \xrightarrow{p} 0 \ .$$

For a finite N, one way to estimate the smooth function $h(x)$ in practice is to use some spline functions based on the estimated $\hat{h}_m(x)$ at m nodes.

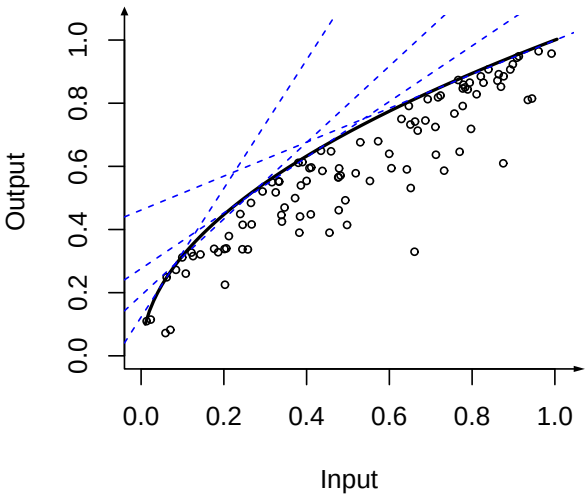

Figure 1: A typical situation : We estimated four tangent lines for the efficient frontiers from simulated 100 data.

We illustrate a typical situation of the present problem as Figure 1. There are 100 firms with a common technology $Y = X^{0.3}$ $(X > 0)$ to produce an output Y and one input X in an economy. Although there could be efficient firms in an industry or a market, but most firms are inefficient and the inefficiency can be denoted as U $(U \leq 0)$, where U is a (non-positive) continuous random variable. We generated a set of random variables from the negative exponential distribution. Since we do not know the exact form of the underlying technology $f(X) = X^{0.3}$ except the fact that Y $(= f(X) + U)$ and f is non-negative and concave, and our task is to estimate the unknown function f nonparametrically from a set of data (X_i, Y_i) $(i = 1, \cdots, 100)$. Then, we try to draw several lines locally by using a set of data around some value at X, which are tangent to the true efficient technology curve at X. We have four estimated tangent lines in Figure 1.

We will propose two non-parametric ways to solve the present statistical problem in this study. In the k-th interval, we set $n = n_k$ $(k = 1, \cdots, m)$ and m is fixed. We consider the problem of estimating the tangent line of $h(X)$ in $\mathrm{I}_k^{(m)}$, and given any $X = x(> 0)$ such as

$$Y_i = a_k + b_k X_i + U_i \ \ (i = 1, \cdots, n)\ . \tag{2.8}$$

We sometimes use notation that $a = a_k, b = b_k$ and $a_k + b_k x \geq h(x)$, $X_i \in \mathrm{I}_k^{(m)} =$

$(w_1^{(k)}, w_2^{(k)}]$ and a_k and b_k are unknown parameters whenever to make no confusion. Then, we will consider the estimation problem a and b in the k-th interval $\mathrm{I}_k^{(m)}$ for a positive integer k. Our proposal is to use the tangent function $a + bx$ to estimate the unknown function $h(x)$ at many points of x.

2.2 The first estimation method

When the sample size of data is not large, we develop the first estimation method, which is based on the linear regression model in each interval. We should note that the first estimation method can be improved substantially when the sample size of data is large, however, as we shall discuss in Section 3.

In the k-th interval, we use the regression slope coefficient

$$\hat{b}_k^{LS} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{2.9}$$

and the maximum of intercept coefficient

$$\hat{a}_k^{LS} = \min_{i=1,\cdots,n} \{a | a + \hat{b}_k X_i \geq Y_i\} \ , \tag{2.10}$$

where $n = n_k$, $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$ and $\bar{X} = (1/n)\sum_{i=1}^n X_i$.

Here, we need the monotonicity restrictions on the estimated coefficients and impose the conditions with k $(k = 1, \cdots, m)$ such that

$$0 \leq \hat{a}_1^{LS} \leq \cdots \leq \hat{a}_m^{LS} \ , \ \hat{b}_1^{LS} \geq \cdots \geq \hat{b}_m^{LS} \geq 0 \ .$$

When the estimated coefficients in an interval do not satisfy the restrictions, we simply disregard the estimated coefficients and the information in the associated intervals. We have the following asymptotic result.

Theorem 1 : Assume that U_i (≤ 0) is a sequence of i.i.d. random variables with the variance $\mathbf{V}[U_i] = \sigma_u^2 < +\infty$, the density $f(u)$ is bounded and smooth at $u = 0$, and X_i are bounded.
(i) Then, in each interval $\mathrm{I}_k^{(m)}$, as $n\ (= n_k) \to \infty$

$$\begin{bmatrix} \hat{a}_k^{LS} - a_k \\ \hat{b}_k^{LS} - b_k \end{bmatrix} \xrightarrow{p} \mathbf{0} \ . \tag{2.11}$$

(ii) As $n\ (= n_k) \to \infty$

$$\sqrt{n}(\hat{b}_k^{LS} - b_k) \xrightarrow{w} N(0, \frac{\sigma_u^2}{M_x}) \ , \tag{2.12}$$

where we assume $M_x = \lim_{n\to\infty}(\frac{1}{n})\sum_{i=1}^n (X_i - \bar{X})^2$ is a positive constant.
(iii) For any $0 < \alpha < 1/2$,

$$n^\alpha(\hat{a}_k^{LS} - a_k) \xrightarrow{p} 0 \ . \tag{2.13}$$

as $n \to \infty$

Proof of Theorem 1 : We use the standard arguments of linear regression in the first part. By using (2.8) and (2.9), we write

$$\sqrt{n}(\hat{b}_k^{LS} - b_k) = \frac{\frac{1}{\sqrt{n}} \sum_{j=1}^n (X_j - \bar{X})(U_j - \bar{U})}{\frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2} \ . \tag{2.14}$$

The denominator converges to M_x and the numerator converges to $N(0, \sigma_u^2 M_x)$ in distribution by applying the central limit theorem (CLT).
Next, we use the relation

$$\begin{aligned} \hat{a}_k^{LS} - a_k &= \min_{i=1,\cdots,n} \{\alpha | Y_i \le a + \hat{b}_k X_i\} - a_k \\ &= \max_{i=1,\cdots,n} \{Y_i - \hat{b}_k X_i\} - a_k \\ &= \max_{i=1,\cdots,n} \{U_i + (b_k - \hat{b}_k) X_i\} \ . \end{aligned}$$

We use the assumption that X_i is bounded ($|X_i| \le X^*$ for some X^*) and $0 < \alpha < 1/2$. Then we can take β such that $0 < \alpha < \beta < 1/2$. For $0 < \alpha < 1/2$,

$$|n^\alpha (b_k - \hat{b}_k) X_i| \le |n^{\frac{1}{2}} (b_k - \hat{b}_k) \max\{X_i\}| [\frac{n^\alpha}{\sqrt{n}}] \xrightarrow{p} 0$$

as $n \to \infty$.
Then for any sequences $z_n = x/n^\alpha$ and $x < 0$,

$$\begin{aligned} P(\max_{i=1,\cdots,n} (U_i + \frac{\epsilon}{n^\beta}) \le z_n) &= \prod_{i=1}^n P(U_i \le \frac{x}{n^\alpha} - \frac{\epsilon}{n^\beta}) \\ &= \exp\{\sum_{i=1}^n \log F(\frac{x}{n^\alpha} - \frac{\epsilon}{n^\beta})\} \ , \end{aligned}$$

where F is the distribution function.
Because U_i is a sequence of i.i.d. random variables with the density f (F is smooth at zero with $F(0) = 1$), the right-hand-side becomes approximately

$$\exp\{n \log[1 + f(0)[\frac{x}{n^\alpha}]\} \sim \exp\{f(0) \frac{x}{n^{\alpha-1}}\} \to 0$$

and $P(n^\alpha[\hat{a}_k^{LS} - a_k] \le x) \to 0$ for any $x < 0$ as $n \to \infty$. By using a similar argument to $P(\max_{i=1,\cdots,n}(U_i - \frac{\epsilon}{n^\beta}) \le z_n)$, we have the result.
(**Q**.**E**.**D**.)

We notice that the order of convergence in $\hat{b}_k$ is $\sqrt{n}$ while the order of convergence in $\hat{a}_k$ may be n^α and $\alpha \ge 1/2$ because of the last part of Theorem 1. It suggests that we may improve the order of convergence in the estimation of slope $\hat{b}_k$. We

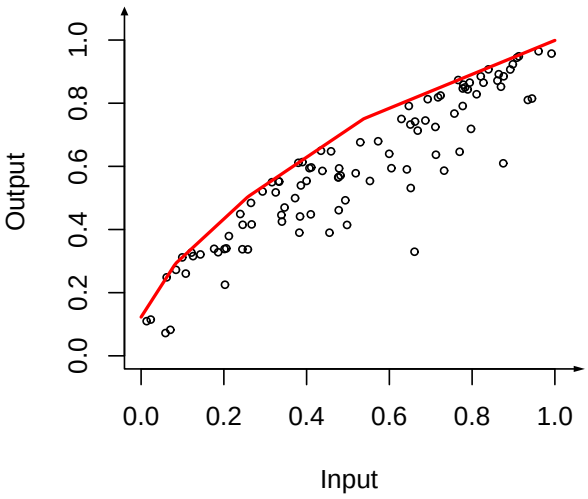

Figure 2: An estimated efficient frontier : For simulated data, we have used the first estimation method to estimate the piece-wise tangent lines.

shall show that the convergence rate is n in the second estimation method.

As a numerical illustration of SDEA by using the first estimation method, we show the estimated efficient frontier in Figure 2 based on some simulated data. Although the true efficient frontier function is continuous and concave in this example, the observed data look non-concave in several intervals because we have a finite number of observations as well as the presence of negative noises. The first estimation method work well because we have used the piece-wise linear efficient frontier functions and we took $m = 5$ in this example. When there are very many observations, the 2nd estimation in the next section may improve the first method and it may have some statistical optimality. However, we need many observations in any fixed interval in the SEVT-based method. (See Section 3.) When the observed data is not large, however, the first estimation method usually gives reasonable solutions for practical purpose in our simulations.

3 The Second Approach of SDEA when the sample size is large

It is possible to improve the first estimation method when the sample size is large. Our second estimation method is based on the statistical extreme value theory (SEVT). The SEVT method has been developed as a branch of statistics, whose focus is on the extremal and rare events such as natural and financial disasters. It has been related to a need to analyze extremal phenomena beyond the standard statistics based on the normal distribution with moments. See Embrechts, P., Klüppelberg, C. and Mikosch (1997) for some details. There are three types of extreme value distributions in SEVT, and we shall use the second (Wiebul) type of extreme distribution because there is an upper bound of observed data in the SDEA models, which is the main target of our statistical analysis.

In this section, we first fix a k ($k = 1, \cdots, m$ and $m \geq 3$) and we order the data $0 \leq w_1^{(1)} \leq X_1 \leq X_2 \leq \cdots \leq X_N \leq w_2^{(m)}$ with $\mathrm{I}_k^{(m)} = (w_1^{(k)}, w_2^{(k)}]$ ($w_1^{(k)} \leq w_2^{(k)}$), $0 \leq w_1^{(1)} < \cdots < w_1^{(m)}$ and $0 \leq w_2^{(1)} < \cdots < w_2^{(m)}$.
We take three consective intervals $\mathrm{I}_j = \mathrm{I}_k^{(m)}(j)$ $j = 1, 2, 3$ and $n(j) = n_k(j)$ with $n = n_k = n(1) + n(2) + n(3)$ ($n(2) \geq 0$) are the numbers of data in each interval. Let $\bar{X}_L = (1/n(1)) \sum_{X_i \in \mathrm{I}_1} X_i$ and $\bar{X}_M = (1/n(3)) \sum_{X_i \in \mathrm{I}_3} X_i$ and we assume that $0 < \bar{X}_L < \bar{X}_M$. Let aslo $Y_L(1) = \max_{X_i \in \mathrm{I}_1} Y_i$ and $Y_M(3) = \max_{X_i \in \mathrm{I}_3} Y_i$.
Then, we define the second estimation method, which is based on the SEVT, by

$$\hat{b}_k = \frac{Y_M(3) - Y_M(1)}{\bar{X}_M - \bar{X}_L} \tag{3.15}$$

and

$$\hat{a}_k = \min_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_2 \cup \mathrm{I}_3} \{a | a + \hat{b}_k X_i \geq Y_i\} \ . \tag{3.16}$$

We impose the monotonicity restrictions on the estimated coefficients in $\mathrm{I}_k^{(m)}$ ($k = 1, \cdots, m$) such that

$$0 \leq \hat{a}_1 \leq \cdots \leq \hat{a}_m \ , \ \hat{b}_1 \geq \cdots \geq \hat{b}_m \geq 0 \ . \tag{3.17}$$

When the estimated coefficients in any interval do not satisfy the necessary restrictions, we simply disregard the estimated coefficients and the associated intervals.

To develop the asymptotic theory, we consider the next condition.
Assume that X_i are bounded and $0 < \bar{X}_L < \bar{X}_M$. For any positive numbers c_1 and c_3 there exists δ (> 0) which satisfies

(**Condition A**) $|X_i - \bar{X}_L| \leq \dfrac{c_1}{n_1^{\delta}}$ for any X_i in I_1 , $|X_i - \bar{X}_M| \leq \dfrac{c_3}{n_3^{\delta}}$ for any X_i in I_3 .

This sufficient condition implies that in two intervals we have a sufficient number of data points in intervals around their means.

For the asymptotic properties of the resulting estimation method, we have the following result on the consistency of $\hat{a}_k$ and $\hat{b}_k$.

Theorem 2 : Assume that U_i (≤ 0) is a sequence of i.i.d. and the ditribution function F has the density $f(u)$ is bounded and smooth at $u = 0$. Also we assume that X_i are bounded X_i are bounded, and $0 < \bar{X}_L < \bar{X}_M$. In (2.8), we consider the case when $n \longrightarrow \infty$ $(n(1), n(3) \to +\infty)$. Then, under Condition A, as $n \to \infty$

$$\begin{bmatrix} \hat{a}_k - a_k \\ \hat{b}_k - b_k \end{bmatrix} \xrightarrow{p} \mathbf{0} \, . \tag{3.18}$$

Proof of Theorem 2 : For z_n and $X_i \in \mathrm{I}_1$,

$$\begin{aligned} P(\max_{X_i \in \mathrm{I}_1} Y_i \leq z_n) &= \prod_{i=1}^{n(1)} P\left(Y_i \leq z_n\right) \\ &= \prod_{i=1}^{n(1)} P\left(Y_i - (a_k + b_k X_i) \leq z_n - (a_k + b_k X_i)\right) \\ &= \prod_{i=1}^{n(1)} P\left(U_i \leq z_n - (a_k + b_k X_i)\right) \, . \end{aligned} \tag{3.19}$$

Since $\delta > 0$, we take α such that $0 < \alpha < \delta$. Then, by taking $z_n = a_k + b_k \bar{X}_l + z/n(1)^\alpha$ $(z < 0)$, The probability can be written as

$$\prod_{i=1}^{n(1)} F\left([\frac{z}{n(1)^\alpha} + b_k(\bar{X}_L - X_i)] \wedge 0 \right) \, . \tag{3.20}$$

Under Condition A, for $\delta > \alpha > 0$ and any $z < 0$, the dominant factor in the right-hand-side becomes

$$\exp[n(1) \log F\left(\frac{z}{n(1)^\alpha}\right)] \to 0$$

as $n(1) \to \infty$.
Hence

$$\max_{X_i \in \mathrm{I}_1} Y_i - (a_k + b_k \bar{X}_L) \xrightarrow{p} 0 \, . \tag{3.21}$$

Then, by using the same argument for I_3 and $\bar{X}_M$,

$$\max_{X_i \in \mathrm{I}_1} Y_i - (a_k + b_k \bar{X}_L) \xrightarrow{p} 0 \, , \max_{X_i \in I_3} Y_i - (a_k + b_k \bar{X}_M) \xrightarrow{p} 0$$

and

$$[\max_{X_i \in \mathrm{I}_2} Y_i - \max_{i \in \mathrm{I}_1} Y_i] - b_k[\bar{X}_M - \bar{X}_L] \xrightarrow{p} 0 \ . \tag{3.22}$$

Hence, we have

$$\hat{b}_k - b_k \xrightarrow{p} 0 \ . \tag{3.23}$$

On the parameter a, we have

$$\begin{aligned} \max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_2 \cup \mathrm{I}_3} [Y_i - \hat{b}_k X_i] &= \max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_2 \cup \mathrm{I}_3} [a_k + b_k X_i + U_i - \hat{b}_k X_i] \\ &= a_k + \max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_2 \cup \mathrm{I}_3} [U_i + (b_k - \hat{b}_k) X_i] \end{aligned} \tag{3.24}$$

and

$$P(\max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_2 \cup \mathrm{I}_3} [Y_i - \hat{b}_k X_i] - a_k \leq z_n) = P(\max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_2 \cup \mathrm{I}_3} [U_i + (b_k - \hat{b}_k) X_i] \leq z_n) \ .$$

Since $b_k - \hat{b}_k \xrightarrow{p} 0$ and X_i are bounded, we can take $\epsilon_n = K/n^{1-\alpha}$ ($\alpha > 0$) such that $P(|(b_k - \hat{b}_k)X_i| \leq \epsilon_n) \to 1$ for a constant K. We take $z_n^* = z_n + \epsilon_n$ (or $z_n = z_n^* - \epsilon_n$) and apply the arguments of the last part of the proof of Theorem 1 to find

$$\hat{a}_k - a_k \xrightarrow{p} 0 \ . \tag{3.25}$$

(**Q.E.D.**)

By constructing the estimated efficiency frontier as

$$\hat{h}_m(x) = \hat{a}_k + \hat{b}_k x \ \ (\text{any } x \in \mathbf{I}_k^{(m)}) \ , \tag{3.26}$$

we have a consistent estimator of the piece-wise function $h(_m x)$, It is because $\hat{h}_m(x) - h_m(x) = (\hat{a}_k - a_k) + (\hat{b}_k - b_k)x \xrightarrow{p} 0$.

For the asymptotic distribution of the estimated coefficients, we need a strong condition with Condition A. The condition implies that we have a sufficient number of data points in two intervals around their means. The asymptotic property of estimator of unknown coefficients depends on the value of δ in Condition A.

Theorem 3 : Assume that U_i (≤ 0) is a sequence of i.i.d. and the distribution function F has the density $f(u)$, which is bounded and smooth at $u = 0$. Also we assume that X_i are bounded X_i are bounded, and $0 < \bar{X}_L < \bar{X}_M$. In (2.8), we consider the case when $n \longrightarrow \infty$ ($n(1), n(3) \to +\infty$).
(i) When $0 < \delta \leq 1$ under Condition A, for any $0 < \alpha < \delta$

$$n^{\alpha}(\hat{b}_k - b_k) \xrightarrow{p} 0 \ . \tag{3.27}$$

as $n \to \infty$.
(ii) When $\delta > 1$ under Condition A, we have the asymptotic distribution of $\hat{b}_k$ as $n \to \infty$,

$$n(\hat{b}_k - b_k) \xrightarrow{w} Z_b = \lambda_3 Z_3 - \lambda_1 Z_1 \ , \tag{3.28}$$

where Z_i $(i = 1, 3)$ follows $G(\lambda) = e^{\lambda z_i}$ $(z_i \leq 0; i = 1, 3)$ and $\lambda = f(0)$. The distribution of Z_b follows $G_b(z) = [\lambda_3/(\lambda_1+\lambda_3)\exp[\frac{\lambda}{\lambda_3}z]$ $(z < 0)$, $G_b(z) = [-\lambda_1/(\lambda_1+\lambda_3)[1 - \exp[\frac{\lambda_3}{\lambda_1}z]$ $(z \geq 0)$, where $\lambda_1 = [1/(\bar{X}_M - \bar{X}_L)][\lim_{n,n(1)\to\infty} \frac{n}{n(1)}]$ and $\lambda_3 = [1/(\bar{X}_M - \bar{X}_L)][\lim_{n,n(3)\to\infty} \frac{n}{n(3)}]$, provided that they converge to finite values for $\lambda_1 > 0$ and $\lambda_3 > 0$.
(iii) For any $0 < \alpha < \min\{\delta, 1\}$,

$$n^\alpha(\hat{a}_k - a_k) \xrightarrow{p} 0 \ . \tag{3.29}$$

as $n \to \infty$

Proof of Theorem 3 : (i) We apply the proof of Theorem 2 for $n^\alpha(\hat{b}_k - b_k)$. For $0 < \alpha < \delta \leq 1$, we take $z_n = a_k + b_k + z/n^\alpha$ $(\alpha < 0)$, then $n^\alpha[\max_{\mathrm{I}_1} Y_i - (a_k + b_k\bar{X}_L] \xrightarrow{p} 0$ and then, we have $n^\alpha(\hat{b}_k - b_k) \xrightarrow{p} 0$.
(ii) Consider the case when $\delta > 1$ under Condition A. For the asymptotic distribution of $\hat{b}_k$, let $Z_{1n} = n(1)[\max_{\mathrm{I}_1} Y_i - (a_k + b_k\bar{X}_L)]$ and $Z_{3n} = n(3)[\max_{\mathrm{I}_3} Y_i - (a_k + b_k\bar{X}_M)]$. Then

$$n(\hat{b}_k - b_k) = \frac{n}{\bar{X}_M - \bar{X}_L}[\frac{Z_{3n}}{n(3)} - \frac{Z_{1n}}{n(1)}] = \lambda_{3n}Z_{3n} - \lambda_{1n}Z_{1n} \ , \tag{3.30}$$

where $\lambda_{1n} = \frac{n}{n(1)(\bar{X}_M - \bar{X}_L)}$ and $\lambda_{3n} = \frac{n}{n(3)(\bar{X}_M - \bar{X}_L)}$.
In the proof of Theorem 2, we set $\alpha = 1$. Then,
$\exp[n(1)\log F(z/n(1))] \to \exp[f(0)z]$ $(z < 0)$ because $F(0) = 1$ and F is smooth at zero with density $f(z)$. Since Z_{1n} and Z_{3n} are independent, the joint asymptotic distribution of Z_{1n} and Z_{3n} is given by

$$G(z_1, z_3) = \exp[\lambda(z_1 + z_3)] \ (z_1 \leq 0, z_3 \leq 0) \ ,$$

where $\lambda_1 = \lim_{n\to\infty} \lambda_{1n}$ and $\lambda_3 = \lim_{n\to\infty} \lambda_{3n}$, We need some care on the asymptotic distribution of $\hat{b}_k$ because $Z_1 \leq 0$ and $Z_3 \leq 0$ and $Z = \lambda_3 Z_3 - \lambda_1 Z_1$ can take positive and negative values. When $Z = \lambda_3 Z_3 - \lambda_1 Z_1 \geq 0$, $\{Z \leq z\}$ and $Z_3 \leq 0$ imply $(\lambda_3 - z)/\lambda_1 \leq Z_1 \leq (\lambda_3/\lambda_1)Z_3$. When $Z = \lambda_3 Z_3 - \lambda_1 Z_1 \leq 0$, $\{Z \leq z\}$ and $Z_1 \leq 0$ imply $Z_3 \leq (\lambda_1 Z_1 + z)/\lambda_3$. Hence we need to consider two cases, separately.
For $z < 0$ is given by

$$\begin{aligned} P(Z \leq z) &= P(Z_3 - \frac{\lambda_1}{\lambda_3}Z_1 \leq \frac{1}{\lambda_3}z) \\ &= \int_{-\infty}^{0} \left[\int_{-\infty}^{(\lambda_1 z_1 + z)/\lambda_3} \lambda^2 \exp \lambda(z_1 + z_3) dz_3 \right] dz_1 \\ &= \frac{\lambda_3}{\lambda_1 + \lambda_3} \exp[\frac{\lambda}{\lambda_3}z] \ . \end{aligned}$$

For $z \geq 0$, we have an evaluation as

$$\begin{aligned} P(Z \leq z) &= \int_{-\infty}^{0} \left[\int_{(\lambda_3 z_3 - z)/\lambda_1}^{(\lambda_3/\lambda_1)z_3} \lambda^2 \exp \lambda(z_1 + z_3) dz_1 \right] dz_3 \\ &= \int_{-\infty}^{0} \lambda \exp(\lambda z_3)[\exp(\lambda(\lambda_3/\lambda_1)z_3) - \exp(\lambda((\lambda_3 z_3 - z)/\lambda_1)z_3)] dz_3 \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_3}[1 - \exp(-\frac{\lambda}{\lambda_1} z)] \ . \end{aligned}$$

(iii) For the asymptotic property of $\hat{a}_k - a_k$, we use the same arguments as the last part of the proof of Theorem 1. Since $n(\hat{b}_k - b_k)$ has a limiting distribution, for $0 < \alpha < 1$ and any negative value z, $P(n^\alpha(\hat{a}_k - a_k) \leq z) \to 0$ as $n \to \infty$, provided that both $n(1)/n$ and $n(3)/n$ converge to positive constants. Then, we have the result in Theorem 3.
(**Q**.**E**.**D**.)

When $n(1) = n(3)$, the distribution of Z_b is the double exponential distribution. It is important to notice that the order of convergence in the asymptotic distribution of $\hat{b}_k$ is n instead of $\sqrt{n}$. It is due to the fact that we use the estimation method based on the maximum value in the intervals.

The asymptotic distribution of $\hat{a}$ is currently not available. However, we may impose an additional condition that there exists a small positive real number ϵ (> 0) such that for any $z < 0$,

$$(\textbf{Condition B}) |P(n[\max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_2 \cup \mathrm{I}_3} (Y_i - \hat{b} X_i) - a_k] \leq z) - P(n[\max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_3} (Y_i - \hat{b} X_i) - a_k] \leq z)| < \epsilon \ .$$

Since we are estimating the piece-wise linear functions, this may not be very restrictive. Some further analysis would be needed.
Then, when $\delta > 1$ with Condition B, we have

$$P(n[\max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_3} (Y_i - \hat{b}_k X_i) - a_k] \leq z) \to H(z) \ , \tag{3.31}$$

where $H(z) = d_1 \exp[e_1 \lambda z]$ for $z \leq 0$ and $H(z) = (1 - d_1) \exp[e_2 \lambda z]$ for $z \leq 0$, where $d_1 = \lim_{n \to \infty} n(2)\bar{X}_M / [n(2)\bar{X}_M + n(1)\bar{X}_L]$, $e_1 = \lim_{n \to \infty} n(1)[\bar{X}_M - \bar{X}_L]/[n\bar{X}_M]$, and $e_2 = \lim_{n \to \infty} n(2)[\bar{X}_M - \bar{X}_L]/[n\bar{X}_L]$, provided that these quantities are well defined and $\bar{X}_M > \bar{X}_L$.
The derivation of $H(z)$ is similar to the first part of the proof of Theorem 3. By using

$$\begin{aligned} & P(n[\max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}_3} (Y_i - \hat{b}_k X_i) - a_k] \leq z) \\ = \ & P(\max\{n[\max_{X_i \in \mathrm{I}_1} (U_i - (b_k - \hat{b}_k) X_i, n[\max_{X_i \in \mathrm{I}_3} (U_i - (b_k - \hat{b}_k) X_i\} \leq z) \ , \end{aligned}$$

under Condition B, it is asymptotically equivalent to

$$\begin{aligned} & P(\max\{n[\max_{X_i\in \mathrm{I}_1}(U_i-(b_k-\hat{b}_k)\bar{X}_L, n[\max_{X_i\in \mathrm{I}_3}(U_i-(b_k-\hat{b}_k)\bar{X}_M\} \le z) \\ \sim\ & P(\max\{c_{11}Z_1+c_{12}Z_2, c_{21}Z_1+c_{22}Z_2\} \le z) \\ =\ & P(c_{11}Z_1+c_{12}Z_2 \le z)\ , \end{aligned}$$

because $c_{11}Z_1+c_{12}Z_2 = c_{21}Z_1+c_{22}Z_2$, where $c_{11} = n/n(1)+\lambda_1\bar{X}_L, c_{12} = -\lambda_2 X_L$, $c_{21} = \lambda_1\bar{X}_M$ and $c_{22} = n/n(2)-\lambda_2\bar{X}_M$.
Let $Z_a^* = c_{11}Z_1+c_{12}Z_2$ and we shall derive its distribution function as follows.
For $z \le 0$, $Z_1 \le (z-c_{12}z_2)/c_{11} \le 0$ $(c_{11} > 0 > c_{12})$,

$$\begin{aligned} H(z) &= \int_{-\infty}^{0}\int_{-\infty}^{(z-c_{12}z_2)/c_{11}} \lambda^2\exp[\lambda(z_1+z_2)]dz_1dz_2 \\ &= \int_{-\infty}^{0}\lambda\exp[\lambda z_2\exp\lambda[(z-c_{12}z_2)/c_{11}]dz_2 \\ &= \frac{c_{11}}{c_{11}-c_{12}}\exp[\frac{\lambda}{c_{11}}z]\ . \end{aligned}$$

For $z > 0$, $z_1 \le 0$ and $(c_{11}z_1-z)/(-c_{12}) \le Z_2 \le 0$, and

$$\begin{aligned} H(z) &= \int_{-\infty}^{0}\int_{(c_{11}z_1-z)/(-c_{12})}^{0} \lambda^2\exp[\lambda(z_1+z_2)]dz_2dz_1 \\ &= \int_{-\infty}^{0}\lambda\exp[\lambda z_1][1-\exp\lambda(c_{11}z_1-z)/(-c_{12})]dz_1 \\ &= 1-\lambda\exp[\frac{\lambda}{c_{12}}z]\int_{-\infty}^{0}\exp[\lambda(1-c_{11}/c_{12})z_1]dz_1 \\ &= 1-\frac{-c_{12}}{c_{11}-c_{12}}\exp[\frac{\lambda}{c_{12}}z]\ . \end{aligned}$$

The distribution function $H(z)$ could be used to approximate the limiting distribution of $\hat{a}$ in practice. Other methods including resampling could be used, but it is beyond the scope of the present study.
For the piece-wise linear function $h_m(x)$, we set $X = x$, and when the limiting random variable of $n(\hat{a}-a)$ is given by Z_a, we have some asymptotic representation. Given $x = \bar{X}$, $\hat{h}_m(x)-h_m(x) \xrightarrow{p} 0$, and the limiting random variable can be represented by

$$n[\hat{h}_m(x)-h_m(x)] \xrightarrow{w} Z_h = Z_a+(\lambda_1Z_1-\lambda_2Z_2)x$$

and Z_a could be approximated by Z_a^*.
The asymptotic distribution depends on the unknown parameter λ (>0). It may be natural to use the residuals $\hat{U}_i = Y_i-\hat{a}_k-\hat{b}_kX_i$ in $I_k^{(m)}$ to estimate λ by $(-1)\hat{\lambda}^{-1} = (1/n)\sum_{i=1}^{n}\hat{U}_i$. Then the confidence interval for λ can be constructed.

4 The Case of Repeated Observations

In this section we discuss the relation between our method in Section 3 and the (classical) statistical extreme value theory (SEVT).
We assume that the inefficiency term is a sequence of i.i.d. random variables with the unknown continuous distribution F. We consider the case when we have repeated observations with a fixed X. We denote X_k $(k = 1, \cdots, m)$ and

$$Y_{kj} = b_k X_k + U_{kj} \ (k = 1, \cdots, m; j = 1, \cdots, n_k) \tag{4.32}$$

where U_{kj} (≤ 0) is a sequence of i.i.d. random variables with the distribution function F and the zero intercept coefficient.

We consider the situation that given $X_k = x$, there are many observations in each intervals and $n_k \to +\infty$ under the assumption that F is smooth at zero. We use

$$\begin{aligned} P(\max_{j=1,\cdots,n_k} Y_{kj} \leq z_n) &= \prod_{j=1}^{n_k} P(U_{kj} \leq z_n - b_k X_k) \\ &= \exp\{\sum_{j=1}^{n_k} \log[1 - \frac{1}{n_k} n_k \bar{F}(z_n - b_k X_k \wedge 0)]\} \\ &\sim \exp\{-\frac{1}{n_k} \sum_{j=1}^{n_k} [n_k \bar{F}(z_n - b_k X_k \wedge 0)]]\} \end{aligned}$$

as $n_k \to \infty$ when we take $z_n = z/n_k + b_k X_k$ and $\bar{F}(x) = 1 - F(x)$. We note that $\bar{F}(x)$ is the right-tail of distribution because U_{kj} are non-positive random variables. Then, by using the Taylor expansion of $\bar{F}(x)$ around $x = 0$ $(\bar{F}(0) = 0)$, as $n_k \to \infty$ $(k = 1, \cdots, m)$

$$P(n_k[\max_{j=1,\cdots,n_k} Y_{kj} - b_k X_k] \leq z) \longrightarrow \exp[f(0)z] \ \ (z \leq 0) \ , \tag{4.33}$$

provided that $f(0)$ is bounded.
Since the limiting distribution is the negative-exponential distribution $F^*(u) = \exp[\lambda u]$ $(u \leq 0)$ with λ (> 0), we have $f(0) = \lambda$.

More generally, it is possible to consider the case when the density function $f(x)$ of the inefficiency terms U_{kj} in (4.32) diverges at $x = 0$. A typical case is the pareto-type distribution when there is finite right endpoint, which has the density around zero $f(x) \sim C(-x)^{\alpha-1}$ $(x < 0, \alpha > 0)$ for some C. One class of distributions has the form that for $y = -x$ (> 0)

$$\bar{F}(-y^{-1}) = y^{-\alpha} L(y) \ , \tag{4.34}$$

where $L(y)$ is a slowly varying function and $\alpha > 0$. (A positive function L on $(0, \infty)$ is slowly varying at ∞ if $\lim_{y\to\infty}[L(ty)/L(y)] = 1$ for $t > 0$. See Page 564 of

Embrechts et al. (1997).) Then, Theorem 3.3.12 of Embrechts, P., Klüppelberg, C. and Mikosch (1997) implies that we can choose $c(n_k)^{-1} = -F^{\leftarrow}(1 - n_k^{-1})$ such that as $n_k \to \infty$

$$P(c(n_k)[\max_{j=1,\cdots,n_k} Y_{kj} - b_k X_k] \leq z) \longrightarrow \exp[-(-z)^{\alpha}] \ \ (z \leq 0) \ , \tag{4.35}$$

where $\alpha > 0$ and $F^{\leftarrow}(t) = \inf\{x | F(x) \geq t\}$ for $0 < t < 1$. This formulation is standard in the statistical extreme value theory (SEVT) and it has been called the maximum domain of attraction demain of the second (Weibull) type of distribution. This asymptotic distribution is known as the 2nd-type (Weibull) extreme value distribution. In this case, however, we need to estimate the scale parameter α in the general case, which may not be a trivial task.

In the SDEA problem, it may be possible to consider the general case of (4.34) with α (> 0) for noise or inefficiency term. However, we did not pursue this line generality in the present work because the assumption of the boundedness of density function at $z = 0$ may be appropriate in most applications. We usually use the DEA method to analyze the effieincy frontier function when there are many inefficient firms and a small number of firms is near to the efficient frontier in a particular industry, for instance.

If we furthr have an intercept term as $Y_{kj} = a_k + b_k X_k + U_{kj}$ ($k = 1, \cdots, m; j = 1, \cdots, n_k$) and a is the intercept term, we denote $V_{kj} = U_{kj} - \mathbf{E}[U_{kj}]$ provided that $\mathbf{E}[U_{kj}]$ exists. Then, we can rewrite $Y_{kj} = (a_k + \gamma) + b_k X_k + V_{kj}$ with $\mathbf{E}[V_{kj}] = 0$. For each k, the statistical model becomes linear regression and the estimation of γ and (a_k, b_k) ($k = 1, \cdots, m$) is possible, but the order of convergence in the regression-based estimation method may be $\sqrt{n}$.

There are some situations when there are many observation in the SDEA problem, but they are not necessarily the same as the statistical model with repeated observations discussed in this section. If we can take $c_n > 0$ such that we have many observations in the region $[x - c_n, x + c_n]$ for some x for instance, it is reasonable to utilize the 2nd estimation method discussed in Section 3.

5 A General Case with Several Explanatory Variables

We consider a generalization of Sections 2 and 3, and let p be the number of explanatory variables. We set $p = 2$ although it is straightforward to consider more general cases with some notational as well as numerical complications.

For $j = 1, 2$, let $\mathrm{I}_{k_j}^{(m_j)} = (w_{j1}^{(k_j)}, w_{j2}^{(k_j)}]$ ($w_{j1}^{(k_j)} \leq w_{j2}^{(k_j)}$), $0 \leq w_{j1}^{(1)} < \cdots < w_{j1}^{(m_j)}$ and $0 \leq w_{j2}^{(1)} < \cdots < w_{j2}^{(m_j)}$. For $\mathbf{k} = (k_1, k_2)'$, $\mathbf{m} = (m_1, m_2)'$ and $\mathrm{I}_{k_j}^{m_j} = (w_{j1}^{k_j}, w_{j2}^{k_j}]$ ($j = 1, 2$), we set the $(k_1, k_2)-$th region by $\mathbf{I}_k^{(m)} = \mathrm{I}_{k_1}^{(m_1)} \times \mathrm{I}_{k_2}^{(m_2)}$. (The number of data is

denoted by $n(p,q) = n_{k_1,k_2}(p,q)$.) We estimate the hyperplanes of the form

$$h_m(\mathbf{X}) = a_k + b_{1k}X_1 + b_{2k}X_2 \tag{5.36}$$

in $\mathbf{X} = (X_1, X_2)' \in \bigcup_k \mathbf{I}_k^{(m)}$ with the concavity restrictions.

Let vectors $\mathbf{x} = (x_1, x_2)'$, $\mathbf{x}(i) = (x_1(i), x_2(i))'$ and $\mathbf{x}(j) = (x_1(j), x_2(j))'$ $(i \neq j)$ be in $\bigcup_k \mathbf{I}_k^{(m)}$ and let non-negative scalars λ_i and λ_j $(i \neq j)$. Then the concavity restrictions imply that

$$h_m(\mathbf{x}) \geq \lambda_i h_m(\mathbf{x}(i)) + \lambda_j h_m(\mathbf{x}(j)) \tag{5.37}$$

for any $\mathbf{x} = \lambda_i \mathbf{x}(i) + \lambda_j \mathbf{x}(j)$ and $\lambda_i + \lambda_j = 1$.
It is straightforward to check these conditions numerically at every estimation, but there may be some complications in their numerical evaluations. As an example, we take $i = 1, j = 2$ and explain the following steps.
(**Step 1**) : First, we estimate a hyperplane $h_1(X_1, X_2) = a(1) + b_1(1)X_1 + b_2(1)X_2$ by using all data with the restrictions $a(1) \geq 0$, $b_1(1) \geq 0$, $b_2(1) \geq 0$ in the region $\mathbf{I}(1) = I_1(1) \times I_2(1)$ ($X_1 \in I_1(1)$ and $X_2 \in I_2(1)$).
(**Step 2**) : Next, we take $\mathbf{I}(1)$ and and some different intervals $\mathbf{I}_k$ $(k = 2, \cdots, m)$ near to $\mathbf{I}(1)$ Then, we estimate hyperplanes $h_1(X_1, X_2) = a(2) + b_1(2)X_1 + b_2(2)X_2$ in each regions locally and check the concavity restrictions and non-negativity of coefficients. If they were not satisfied, we disregard the estimation results. If they were satisfied, we use the estimation result and use the piece-wise linear functions.
(**Step 3**) : We repeat the same procedure. Since the number of data is finite, we will stop this procedure eventually. (In our experiments, we have taken m such that m_1 and m_2 are less than $0.1\times$ (sample size).)

5.1 The first estimation method of coefficients

We extend the estimation of unknown coefficients with one explanatory variable to the one with several variables. We illustrate this problem and consider the case of two explanatory variables. We first fix k_1 and k_2, and we take the number of data as $n_{k_1,k_2}(p,q)$ $(p,q = 1,2)$. We apply the least squares estimators of the coefficient vector and construct the intercept coefficient by adjusting the level of output. Then we continue to construct coefficients such that they satisfy the monotonicity restrictions.
For the first regression-based method of coefficients in Section 2, it is straightforward to extend the method in Section 2 to the case when there are several explanatory variables. The coefficients b_{jk} $(j = 1, \cdots, p; k = 1, \cdots, m)$ can be estimated by the linear regression equation

$$\hat{\mathbf{b}}_k^{LS} = [\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^{`}]^{-1} [\sum_{i=1}^{n} \mathbf{X}_i Y_i] \ , \tag{5.38}$$

where $\mathbf{X}_i = (X_{ji})$ is a $p \times 1$ vector of input variables and Y_i is the output variable. The estimator of the intercept coefficient a_k is given by

$$\hat{a}_k^{LS} = \min_{i=1,\cdots,n} \{a | a + \hat{b}_k^{LS} \mathbf{X}_i \geq Y_i\} \ . \tag{5.39}$$

The order of the asymptotic distribution of b_{jk} and a_k $(j = 1, \cdots, p; k = 1, \cdots, m)$ are $\sqrt{n}$ and n, respectively. It is because Theorem 1 and its proof can be extended directly to this case.

It may be straightforward to extend our analysis in this section to the general case when $p \geq 2$ such that for $k = 1, \cdots, m$,

$$Y_i = a_k + \sum_{j=1}^{p} b_{jk} X_{ji} + U_i \ \ (i = 1, \cdots, n) \ , \tag{5.40}$$

where $U_i \ \leq 0$.

5.2 The second estimation method of coefficients

We also extend the 2nd estimation method with the concavity restrictions based on the SEVT method explained in Section 3. As for an illustration, we take the case of $p = 2$, and from $\mathbf{I}_k^{(m)}$, $k = 1, \cdots, m$ we take conswctive 9 regions in the form of

$$\mathbf{I}(i,j) = I_1(i) \times I_2(j) = (w_{1i}, w_{1,i+1}] \times (w_{2j}, w_{2,j+1}] \ \ (i,j = 1,2,3)$$

and $n(i,j)$ denotes the number of data in $\mathbf{I}(i,j)$. Since there are many intervals, in practice we can take empty regions as $n(i,j) = 0$ (i or j=2) in practice. We take the means in each regions as $X_1(i,j)$ and $X_2(i,j)$ $(i,j = 1,2,3)$.

For estimation, we take a combination of j and k and set
$X_1(j,k) = (1/n(j,k)) \sum_{X_i \in I(j.k) X_{1i}}$ $(j,k = 1,3)$ and
$X_2(j,k) = (1/n(j,k)) \sum_{X_i \in I(j.k) X_{2i}}$ $(j,k = 1,3)$,
where $n(j,k)$ are the number of observations in $I_k^{(m)}(j,k)$ $(j,k = 1,2,3)$. The corresponding maximum output value in each regions as $Y_M(j,k) = \max_{X_i \in \mathrm{I}(j,k)} Y_i$ $(j,k = 1,3)$. Then, the following derivations are the direct extensions of Section 3. By using the assumption of smooth distribution function around zero, we first use the relation

$$\begin{aligned} P(\max_{I(3,1)} Y_i \leq z_n) &= P(\max_{I(3,1)} [U_i + a + b_1 X_{1i} + b_2 X_{2i}] \leq z_n) \\ &= \prod_{i=1}^{n(3,1)} P(U_i + a + b_1 X_{1i} + b_2 X_{2i} \leq z_n) \ . \end{aligned}$$

Then, by using the same arguments in section 3, we assume that for any positive number c there exists δ (> 0) such that the sequences $\mathbf{X}_i$ in $\mathbf{I}_k^{(m)}$ satisfy

(**Condition A***) $\qquad \|\mathbf{X}_i - \bar{X}(i,j)\| \leq \dfrac{c}{n^\delta}$ for any $\mathbf{X}_i$ in $\mathbf{I}_(i,j)$ $(i,j = 1 \text{ or } 3)$,

Then, we have

$$\max_{I(3,1)} Y_i - [a + b_1 X_1(3,1) + b_2 X_2(3,1)] \xrightarrow{p} 0 \ .$$

Similarly, we find that $\max_{\mathrm{I}(1,1)} Y_i - [a + b_1 X_1(1,1) + b_2 X_2(1,1)] \xrightarrow{p} 0$ and $\max_{\mathrm{I}(1,3)} Y_i - [a + b_1 X_1(1,3) + b_2 X_2(1,3)] \xrightarrow{p} 0$.
By using the above relations,

$$[Y_M(3,1) - Y_M(1,1)] - b_1[X_1(3,1) - X_1(1,1)] - b_2[X_2(3,1) - X_2(1,1)] \xrightarrow{p} 0$$

and

$$[Y_M(1,3) - Y_M(1,1)] - b_1[X_1(1,3) - X_1(1,1)] - b_2[X_2(1,3) - X_2(1,1)] \xrightarrow{p} 0 \ .$$

We define the estimator $(\hat{b}_1, \hat{b}_2)$ of slope coefficients by

$$\begin{bmatrix} Y_M(3,1) - Y_M(1,1) \\ Y_M(1,3) - Y_M(1,1) \end{bmatrix} = \begin{bmatrix} X_1(3,1) - X_1(1,1) & X_2(3,1) - X_2(1,1) \\ X_1(1,3) - X_1(1,1) & X_2(1,3) - X_2(1,1) \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \ .$$

The estimator $\hat{a}$ of intercept coefficient is defined by

$$\hat{a} = \min_{X_i \in \bigcup_{j,k=1,3} \mathrm{I}(j,k)} \{a | a + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} \geq Y_i\} \ . \tag{5.41}$$

Then, under the assumption that there exists δ (> 0) with Condition A^*, we find that

$$\begin{bmatrix} Y_M(3,1) - Y_M(1,1) \\ Y_M(1,3) - Y_M(1,1) \end{bmatrix} - \begin{bmatrix} X_1(3,1) - X_1(1,1) & X_2(3,1) - X_2(1,1) \\ X_1(1,3) - X_1(1,1) & X_2(1,3) - X_2(1,1) \end{bmatrix} \begin{bmatrix} \hat{b}_1 - b_1 \\ \hat{b}_2 - b_2 \end{bmatrix}$$
$$\xrightarrow{p} \ 0 \ .$$

If we further assume that

$$\mathrm{rank} \begin{bmatrix} 1 & X_1(3,1) & X_2(3,1) \\ 1 & X_1(1,3) & X_2(1,3) \\ 1 & X_1(1,1) & X_2(1,1) \end{bmatrix} = 3 \tag{5.42}$$

for instance, then, $\hat{b}_1 - b_1 \xrightarrow{p} 0$ and $\hat{b}_2 - b_2 \xrightarrow{p} 0$.
Let $Z_n(3,1) = n(3,1)[\max_{I(3,1)} Y_i - (a + b_1 X_1(3,1) + b_2 X_2(3,1))]$,
$Z_n(1,3) = n(1,3)[\max_{I(1,3)} Y_i - (a + b_1 X_1(1,3) + b_2 X_2(1,3))]$,
and $Z_n(1,1) = n(1,1)[\max_{I(1,1)} Y_i - (a + b_1 X_1(1,1) + b_2 X_2(1,1))]$.
Then, under the assumption that there exists δ (> 1) with Condition A^*, we have the limiting exponential random variables $Z(3,1)$, $Z_(1,3)$, and $Z(1,1)$ with the joint distribution

$$G(z_{31}, z_{13}, z_{11}) = \exp[\lambda(z_{31} + z_{13} + z_{11})] \ \ (z_{31} \leq 0, z_{13} \leq 0, z_{11} \leq 0) \ .$$

Then, the asymptotic distribution of $n[\hat{b}_1 - b_1, \hat{b}_2 - b_2]$ is the weighted average of exponential distribution in the expression

$$\mathbf{Z}_b = \begin{bmatrix} X_1(3,1) - X_1(1,1) & X_2(3,1) - X_2(1,1) \\ X_1(1,3) - X_1(1,1) & X_2(1,3) - X_2(1,1) \end{bmatrix}^{-1} \begin{bmatrix} \lambda(3,1) & 0 & -\lambda(1,1) \\ 0 & \lambda(1,3) & -\lambda(1.1) \end{bmatrix} \begin{bmatrix} Z(3,1) \\ Z(1,3) \\ Z(1,1) \end{bmatrix} ,$$

where $\lambda(3,1) = \lim_{n\to\infty} n/n(3,1)$, $\lambda(1,3) = \lim_{n\to\infty} n/n(1,3)$, and
$\lambda(1,3) = \lim_{n\to\infty} n/n(1,1)$ as $n, n(3,1), n(1,3), n(1,1) \to \infty$.
Under the same setting with $\mathbf{x} = \bar{\mathbf{X}}$, $\hat{h}_m(\mathbf{x}) - h_m(\mathbf{x}) \xrightarrow{p} 0$, and the asymptotic distribution of the estimated hyper-planes is given by

$$n[\hat{h}_m(\mathbf{x}) - h_m(\mathbf{x})] \xrightarrow{p} Z_h = Z_a + \mathbf{Z}_b^{'}\mathbf{x} , \tag{5.43}$$

where $\mathbf{x} = (x_1, x_2)^{'}$.
This expression is a direct generalization to the cases when $p \geq 2$.

When the sample size is not large while the number of explanatory variables p is greater than 1, the number of data in each cell may be small. Then the estimation procedure may not be easily used. To avoid this problem, one may use a different procedure to use multi-dimension cells. To illustrate an alternative method, we use the case when $p = 2$, for instance, and we denote each cell as $\mathbf{I}(j,k)$ $(j,k = 1,2,3)$. We also use the notations such that for $j,k = 1,2,3$, $\bigcup_k \mathbf{I}(j,k) = \mathbf{I}(j,\cdot)$ and $\bigcup_j \mathbf{I}(j,k) = I(\cdot,k)$.
To cope with this problem, first, as we have mentioned, we can take null regions $\mathbf{I}(i,j)$ when i or j is 2. Second, we use the relation

$$P(\max_{\bigcup_k \mathrm{I}(3,k)} Y_i \leq z_n) = P(\max_{\bigcup_k \mathrm{I}(3,k)} [U_i + a + b_1 X_{1i} + b_2 X_{2i}] \leq z_n) .$$

Then we can develop the similar evaluation except the fact that the resulting limit random variables $Z_n(j,\cdots)$ and $Z_n(\cdot,k)$ $(j,k = 1,3)$ are correlated even when $n \to \infty$. The limiting distributions of estimators of coefficients can be expressed by the limiting joint random variables $Z(j,\cdots)$ and $Z(\cdot,k)$ $(j,k = 1,2)$, which follow

$$G(z_{j,\cdot}, z_{\cdot,k}) = \exp\{\lambda \sum_{j,k} [z_{j,\cdot} \wedge z_{\cdot,k}]\} \ (j,k = 1,3), \tag{5.44}$$

where $z_{j,\cdot,\cdot} \leq 0, z_{\cdot,k,\cdot} \leq 0$, $\lambda(j,k) \sim n/n(j,k)$.
Then, the asymptotic distribution of $n[\hat{b}_1 - b_1, \hat{b}_2 - b_2]$ is the weighted average of exponential distribution in the expression of

$$\mathbf{Z}_b^* = \begin{bmatrix} X_1(3,\cdot) - X_1(1,\cdot) & X_2(3,\cdot) - X_2(1,\cdot) \\ X_1(\cdot,3) - X_1(\cdot,1) & X_2(\cdot,3) - X_2(\cdot,1) \end{bmatrix}^{-1} \tag{5.45}$$

$$\times \begin{bmatrix} \lambda(3,\cdot) & -\lambda(1,\cdot) & 0 & 0 \\ 0 & 0 & \lambda(\cdot,3) & -\lambda(\cdot,1) \end{bmatrix} \begin{bmatrix} Z(3,\cdot) \\ Z(1,\cdot) \\ Z(\cdot,3) \\ Z(\cdot,1) \end{bmatrix}.$$

This representation can be extended straightforwardly to the cases when $p \geq 3$. The detail of this procedure is currently investigation, but it seems that we need a simulation-based evaluation of the limiting distribution.

6 Efficient Frontier and Measurement Errors

There are cases when we should not ignore the measurement errors in inputs and outputs in the SDEA problem. Let V_i be the measurement errors for the i-th observation. First, when $V_i < 0$, it may not be possible to distinguish it from the inefficiency term, which does not take any positive value. Second, we consider the case when $V_i \geq 0$. A typical case would be $V_i = cf(X_i)^*$, where $f(X_i)^*$ is the *hidden efficient frontier* and c is a non-negative measurement error rate. Then we have the statistical model (2.1) as $Y_i = f(X_i) + U_i$, where

$$f(X_i) = f(X_i)^*(1+c)\ , \tag{6.46}$$

Then, it may be reasonable to to estimate the frontier function without measurement errors by $\hat{f}(X_i)^* = \hat{f}(X_i)/(1+c)$. There can be some examples of reporting inaccurate numbers and accounting misconducts as typical examples of positive measurement errors. In such cases, their roles could not be ignorable.

However, when we have measurement-errors in the SDEA problem, there is an alternative approach to treat them as outliers. One empirical example will be reported in the next section.

7 An Empirical Example : Life-Insurance Industry in Japan

As an empirical example, we have applied the SDEA method in the previous sections to the accounting data sets on the life-insurance industry in Japan, which are public data during 2017-2021 fiscal years in "Seimei-Hoken-Jigyou Gaikyou" (Seimei-Hoken-Kyokai (2021)).

We have used the data as (1) works : office workers, (2) capital : total shareholders' equity, (3) expense : operating expenses, (4) insurance : total payment of insurance benefits, and (5) income : ordinary income. The output variable is the ordinary income.

Since there are 41 companies in this industry, which is rather small, we have used the first method to estimate the efficient frontier function. Among 41, there is one firm, Kanpo-Seimei, which is quite different from others because of the long-history and some institutional changes. Then, we may need to exclude this firm to estimate the efficiency frontier. Apparently, the monotonicity and concavity assumption on the efficient frontier is not satisfied as we illustrated the problem in Figure 3. Thus, it is appropriate to treat Kanpo-Seimei as an outlier and should be deleted, which is not discussed in detail, but we briefly mention to the fact that the historical role of the life-insurance industry has quite different from other industrialized countries like U.K. and U.S.. There were some historical as well as institutional reasons why there are a few major life-insurance companies in Japan and the number of life-insurance companies is small in comparison with those in the U.S. and U.K.. Kanpo-Seimei was originally a part of the National Post Office in Japan, and it was privatized in 2006, for instance. See Kubo (2011) for some details of the historical development of the life and non-life insurance industries in Japan.

In our analysis we have focused on the data on 40 companies in our empirical analysis. We used the number of office workers as an input and ordinary income as the output and estimated the 2021 efficient frontier in Figure 4. We also used Capital as an input and ordinary income as the output and estimated the 2021 efficient frontier in Figure 5. From these two figures we have found that we can estimate the frontiers in a reasonable manner. That is, there are several companies, which are close to the efficient frontier and there are other inefficient companies. We also found that there are only several large companies in the life-insurance industry, the estimation of the efficient frontier in the right-hand area is statistically a difficult problem.

8 Concluding Remarks

In this paper, we discussed the problem of DEA and have developed a new SDEA method based on the statistical modeling of linear regression and extreme value distribution, which may be new to both the operations research and statistics communities. We also report an empirical analysis of life-insurance industry in Japan as an application. Because the number of data is quite small in our example, we used the linear regression based method for estimating coefficients. When the number of data is large, however, we have shown that some efficiency gain in the statistical estimation could be obtained under some additional conditions if we use the statistical extreme value (SEVT) method.

There are a number of problems in the SDEA method remained to be investigated. First, we could not have obtained the asymptotic distribution of the intercept parameter a in two estimation methods. Second, the asymptotic theory of the second estimation method without Conditions A, B, and/or C is currently an important

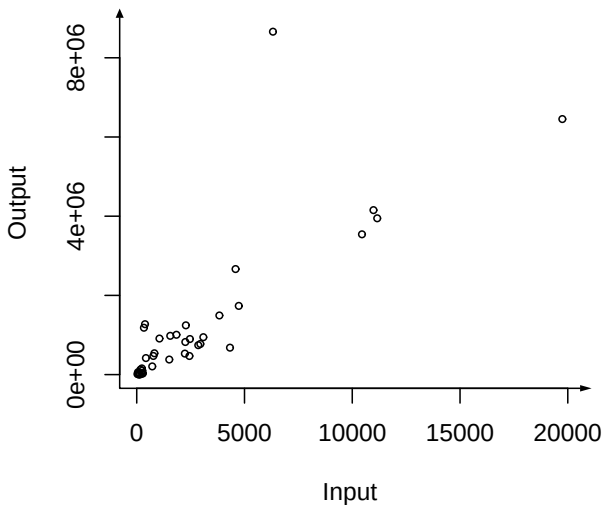

Figure 3: An outlier situation : In the life-insurance industry in Japan, there is an outlier and there are some reasons why it is.

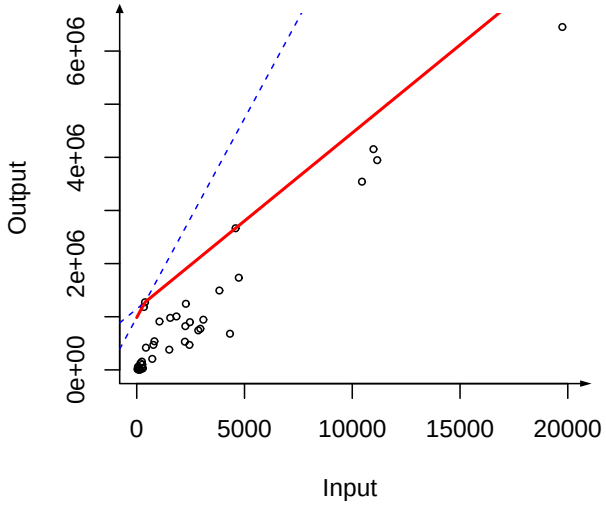

Figure 4: An estimated frontier : Input is Workers and output is Income.

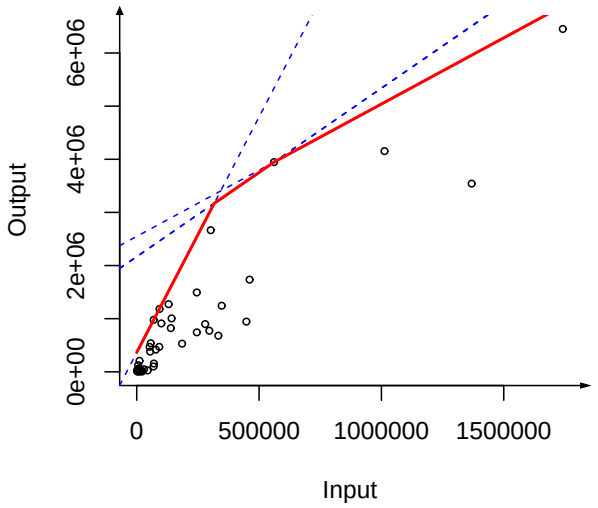

Figure 5: An estimated frontier : Input is Capital and output is Income.

research topic. Third, the statistical models treated in this paper can be generalized to several directions including multivariate inputs and outputs. In some cases, however, it is not a trivial task to impose the monotonicity and concavity restrictions when we estimate the estimated frontiers from a finite set of data. If we had a huge number of data, it may be possible to use explanatory variables in an efficient way.

Another important statistical issue would be that there can be several procedures to choose the number of intervals (m) in a finite number of data analysis and we need to develop some criterion of selecting the number of interval nodes (m) in some optimal way given a finite number of data. It includes some methods to choose intervals randomly.

We are currently investigating various aspects of theoretical problems and applications of the SDEA method proposed in the present work. We are also developing the R-programs for numerical evaluations.

References

[1] Aigner, D., K. Lovell, and P. Schmidt (1977), "Formulation and Estimation of Stochastic Production Models," *Journal of Econometrics*, 6, 21-37.

[2] Cooper, W. W., Seiford, L. M., and Tone, K. (2007), *Data envelopment analy-*

sis: a comprehensive text with models, applications, references and DEA-solver software, 2nd edition, New York: Springer.

[3] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997), *Modelling Extremal Events*, Springer.

[4] Green, W.H. (2003), *Econometric Analysis*, Prentice Hall.

[5] Kubo, H. (2011), "Measuring the Effects of Management Integration in Insurance Industries of Japan" (in Japanese), Hokengaku-Zatsushi (the Journal of Insurance Science), Hoken-Gakkai (The Japanese Society of Insurance Science).

[6] Seimei-Hoken-Jigyou Gaikyou (2021), (in Japanese), Seimei-Hoken-Kyokai (Life Insurance Association in Japan).