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

A Statistical Data Envelopment Analysis *

Naoto Kunitomo †
and
Yu Zhao ‡

November 19, 2022

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 literatures as well as statistical community. We first consider the basic statistical DEA model, in which the observed data is the sum of an increasing concave function of inputs and a random noise (or inefficiency) term taking only non-positive value. 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 nonparametric statistical analysis, the linear regression analysis and the statistical extreme value theory. 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 2022-11-19. This is a preliminary memorandum. We thank Yasushi Yoshida for a comment on the life-insurance industry in Japan.

†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, which are often based on the mathematical programming techniques. In economics, on the other hand, the parametric statistical estimation method of production frontiers 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.) The main purposes of these two methods are similar, but their traditional approaches and mathematical techniques to solve the problems are quite different.

In this paper, we develop a statistical data envelopment analysis (SDEA), which seems to be new to operations research literatures as well as econometrics and statistical sciences. We first consider the basic statistical DEA model, in which the observed data is the sum of an increasing concave function and a random noise (or inefficiency) term, taking non-positive values. 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 is to use the non-parametric statistical analysis, the regression analysis, and the statistical extreme value theory (SEVT) as the statistical methods 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. Then, 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. 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. We also consider the case when we have measurement errors as well as inefficiencies in the observed data sets.

The main purpose of this paper is to introduce our new statistical approach to the EDA problem and some theoretical results. Then we shall 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 in this case. Since our approach is not along the traditional approaches 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 the first estimation method in the simple case. In Section 3, we give the second estimation method for the case when the sample size is large. In Section 4, we discuss the relation of our SDEA model and its relation

to 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 frontiers. In Section 7, we report an empirical study of the SDEA method for the life-insurance industry in Japan. In Section 8, we provide some concluding remarks.

2. A New Approach of SDEA

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 are non-negative. We assume that the efficient frontier function $h(\cdot)$ may be smooth and twice-differentiable with $h^{'} > 0$ and $h^{''} < 0$. (We often consider the case when we only know that $f(\cdot)$ is a concave function.) Let also the random variable U representing the inefficiency term from the efficient frontier function, and we assume the relation

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

In the standard EDA, 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

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

where $I_k^{(m)} = (w_1^{(k)}, w_2^{(k)}]$ $(\ w_1^{(k)} \leq w_2^{(k)})$, $0 \leq w_1^{(0)} < w_1^{(1)} < \cdots < w_1^{(m)}$ and $0 \leq w_2^{(0)} < w_2^{(1)} < \cdots < w_2^{(m)}$.
In this study, we restrict our foumulation to the case when X_i is a bounded deterministic variable and $w_1^{(0)} \leq X_1 \leq X_2 \leq \cdots \leq X_N \leq w_2^{(m)}$. 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 numbers and as a typical case we take that U_i follows the negative exponential distribution such that for some positive $\lambda > 0$

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

and the basic 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 literatures 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 $I_k^{(m)}$ such that $\bigcup_{k=1}^m I_k^{(m)} = (w_1^{(1)}, w_2^{(m)}]$ and we denote n_k as the number of data in $I_k^{(m)} = (w_1^{(k)}, w_2^{(k)}]$ with

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

We shall investigate some statistical estimation method of the piece-wise linear function $\hat{h}_m$ such that as m is large and as $m \to +\infty$.

$$\sup_x |\hat{h}_m(x) - h(x)| \xrightarrow{p} 0 \ . \tag{2.6}$$

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.

We illustrate a typical situation of the present problem as Figure 1. There are 200 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 the 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) $(i = 1, \cdots, 200)$. 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 that value of X. We have six estimated tangent lines in Figure 1.

We will propose two non-parametric statistical ways to solve the present statistical problem. In the k-th interval, we set $n = n_k$ $(k = 1, \cdots, m)$ and m is fixed.

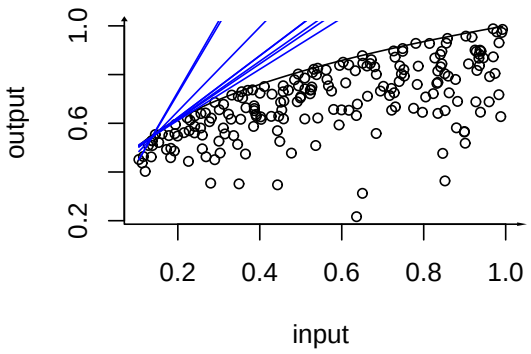

Figure 1: A typical situation : We estimated six tangent lines for the efficient frontiers from simulated data.

We consider the problem of estimating the tangent line of $h(X)$ in $I_k^{(m)}$ any given $X = x(> 0) = (w_1^{(k)} + w_2^{(k)})/2$ such as

$$Y_i = a_k + b_k X_i + U_i \ \ (i = 1, \cdots, n) \ , \tag{2.7}$$

where we often use notation that $a = a_k, b = b_k$ and $a_k + b_k x \geq h(x)$, $X_i \in I_k^{(m)} = (w_1^{(k)}, w_2^{(k)}]$ and a_k and b_k are unknown parameters. We assume that we have $n_k(1)$ observations in $(w_1^{(k)}, x]$, $n_k(2)$ observations in $(x, w_2^{(k)}]$ $(n_k = n_k(1) + n_k(2))$ and $x = (1/n_k) \sum_{i=1}^{n} X_i$ [1].

In the following analysis, we first fix a k $(k = 1, \cdots, m))$ and set $n = n_k$, $n(1) = n_k(1)$, and $n(2) = n_k(2)$ in this subsection because we will consider the estimation problem a and b in the k-th interval $I_k^{(m)}$ for a positive integer k. Then our proposal is to use the tangent function $a + bx$ to estimate the unknown function $h(x)$.

The first estimation method of a and b :

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

[1]It is possible to use the sample median value of X instead of the sample mean.

the first estimation method can be substantially improved 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 coefficient

$$\hat{b}_k^{LS} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{2.8}$$

and

$$\hat{a}_k^{LS} = \min_{i=1,\cdots,n} \{a | a + \hat{b}_k X_i \geq Y_i\} \ , \tag{2.9}$$

where $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$ 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 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 $\mathbf{E}[U_i] = \gamma$, $\mathbf{V}[U_i] = \sigma_u^2 < +\infty$, the density $f(u)$ is bounded and smooth at $u = 0$, and X_i are bounded.
(i) Then, in each interval $I_k^{(m)}$, as $n \to \infty$

$$\begin{bmatrix} \hat{a}_k^{LS} - a_k \\ \hat{b}_k^{LS} - b_k \end{bmatrix} \xrightarrow{p} \mathbf{0} \ . \tag{2.10}$$

(ii) As $n \to \infty$

$$\sqrt{n}(\hat{b}_k^{LS} - b_k) \xrightarrow{w} N(0, \frac{\sigma_u^2}{M_x}) \ , \tag{2.11}$$

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.
As $n \to \infty$

$$n(\hat{a}_k^{LS} - a_k) \xrightarrow{w} Z_a \ , \tag{2.12}$$

where Z_a follows $G_a(z) = \exp[f(0)z_a]$ $(z_a \leq 0)$.

Proof of Theorem 1 : We use the standard arguments of linear regression in the first part. By using (2.7) and (2.8), we write

$$\sqrt{n}(\hat{b}_k^{LS} - b_k) = \frac{\frac{1}{\sqrt{n}}\sum_{j=1}^n (X_j - \bar{X})(U_j - \bar{U})}{\frac{1}{n}\sum_{j=1}^n (X_j - \bar{X})^2} \ . \tag{2.13}$$

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

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

By our assumption, X_i is bounded ($|X_i| \le K$) and then for any positive $\epsilon_n \to 0$, we know that $P(|(\hat{b}_k^{LS} - b_k)X_i| \le \epsilon_n) \to 1$. Then for any positive sequences z_n,

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

where U_i is a sequence of i.i.d. random variables with F. We re-write the last term can be re-written as

$$\exp\{\sum_{i=1}^{n} \log[1 - \frac{1}{n}(n\bar{F}(z_n - \epsilon))]\}$$

where $\bar{F}(x) = 1 - F(x)$ (i.e. it is the tail probability because $F(0) = 1$). By using the Talor expansion of $\bar{F}(z)$ around $z = 0$ and setting $\lim_{n\to\infty} n(z_n - \epsilon_n) = z$, we can approximate $-n\bar{F}(z_n - \epsilon_n) \sim f(0)(z_n - \epsilon_n) \to f(0)z$ as $n \to \infty$. Similarly, by evaluating $P(\max_{i=1,\cdots,n}(U_i - \epsilon_n) \le z_n))$ and setting $\lim_{n\to\infty} n(z_n + \epsilon_n) = z$, we obtain the same limiting distribution. Since the difference of $P(n[\hat{a}_k^{LS} - a_k] \le z)$ and $P(\max_{i=1,\cdots,n}(U_i + \epsilon_n) \le z_n$ with $n(z_n - \epsilon_n) = z$ is stochstically negligible as $n \to \infty$, 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 $\hat{a}_k$ is n because of the structure of our problem. It suggests that we may improve the order of convergence in the estimation of slope $\hat{b}_k$. We 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 a continuous 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

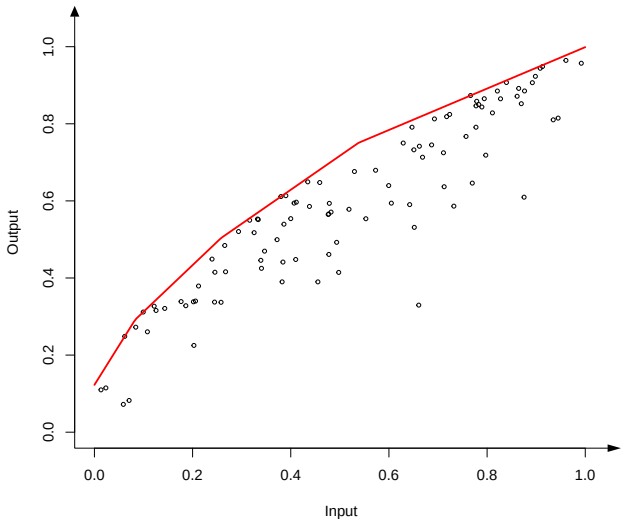

Figure 2: An estimated efficient frontier : For simulated data, we used the second estimation method to estimate the piece-wise tangent lines.

functions and we took $m = 5$ in this example. When there are very many observations, the 2nd estimation in the next section work well, which has some statistical optimality. When there are many observations in any fixed intervals, the SEVT method works in each intervals. (See Sections 3 and 4 for the detail.) When the observed data is not large, however, the first estimation method usually give reasonable solutions for practical purposes 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 disasters. 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 type of extreme distribution because of the SDEA structure that there is an upper bound of observed data, which is the main target of the SDEA problem in this paper.

In this section, we also first fix a k $(k = 1, \cdots, m)$ and we order the data as $0 < X_1 \leq \cdots \leq X_n$ in $I_k^{(m)}$. Let $X_L = (1/n(1)) \sum_{i=1}^{n(1)} X_i$ and $X_M = (1/n(2)) \sum_{i=n(1)+1}^{n} X_i$

$(n = n(1) + n(2))$, and we assume that $0 < X_L < X_M$. Let aslo $Y_M(1) = \max_{1\le i\le n(1)} Y_i$ and $Y_M(2) = \max_{n(1)+1\le i\le n} Y_i$.
Then we define the second estimation method, which is based on the SEVT, by

$$\hat{b}_k = \frac{Y_M(2) - Y_M(1)}{X_M - X_L} \tag{3.1}$$

and

$$\hat{a}_k = \min_{i=1,\cdots,n} \{a | a + \hat{b}_k X_i \ge Y_i\} \ . \tag{3.2}$$

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

$$0 \le \hat{a}_1 \le \cdots \le \hat{a}_m \ , \ \hat{b}_1 \ge \cdots \ge \hat{b}_m \ge 0 \ . \tag{3.3}$$

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

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. negative-exponential random variables with λ (> 0) and X_i are bounded. In the present model, we consider the case when $n \longrightarrow \infty$ $(n(1), n(2) \to +\infty)$. Then, as $n \to \infty$

$$\begin{bmatrix} \hat{a}_k - a_k \\ \hat{b}_k - b_k \end{bmatrix} \xrightarrow{p} \mathbf{0} \ . \tag{3.4}$$

Proof of Theorem 2 : Let Y_i $(i = 1, \cdots, n(1))$ and $X_i \in [X_1, X_{n(1)}]$ $(i = 1, \cdots, n(1))$. Then

$$\begin{aligned} P(\max_{1\le i\le n} Y_i \le z_n) &= \prod_{i=1}^{n(1)} P\left(Y_i \le z_n\right) \\ &= \prod_{i=1}^{n(1)} P\left(Y_i - (a_k + b_k X_i) \le z_n - (a_k + b_k X_i)\right) \\ &= \prod_{i=1}^{n(1)} P\left(U_i \le z_n - (a_k + b_k X_i)\right) \\ &= \exp[\lambda(n(1) z_n - n_1 a_k - b_k \sum_{i=1}^{n(1)} X_i)] \\ &= \exp[\lambda n(1)(z_n - a_k - b_k \bar{X}_L)] \ , \end{aligned} \tag{3.5}$$

where $\bar{X}_L = (1/n(1)) \sum_{i=1}^{n(1)} X_i$.
If we take $z_n - a_k - b_k \bar{X}_L = -\delta$ ($\delta > 0$), then as $n \to \infty$ the probability becomes 0. If we set $z_n - a_k - b_k \bar{X}_L = 0$, then it becomes 1 as $n \to \infty$. Hence

$$\max_{1 \le i \le n} Y_i - (a_k + b_k \bar{X}_L) \xrightarrow{p} 0 \ . \tag{3.6}$$

We set the data intervals as $I_1 = (X_1, X_{n(1)}]$ and $I_2 = [X_{n(1)+1}, X_n]$ with $n = n(1) + n(2)$. b by $\hat{b}$. Then, by using the same augument on $X_i \in [X_{n(1)+1}, X_{n(1)+n(2)}]$ as $X_i \in [X_1, X_{n(1)}]$,

$$\max_{i \in I_1} Y_i - (a_k + b_k \bar{X}_L) \xrightarrow{p} 0 \ , \max_{i \in I_2} Y_i - (a_k + b_k \bar{X}_M) \xrightarrow{p} 0$$

and

$$[\max_{i \in I_2} Y_i - \max_{i \in I_1} Y_i] - b_k [\bar{X}_M - \bar{X}_L] \xrightarrow{p} 0 \ . \tag{3.7}$$

Hence, we have

$$\hat{b}_k - b_k \xrightarrow{p} 0 \ . \tag{3.8}$$

On the parameter a, we have

$$\begin{aligned} \max_{1 \le i \le n} [Y_i - \hat{b}_k X_i] &= \max_{1 \le i \le n} [a_k + b_k X_i + U_i - \hat{b}_k X_i] \\ &= a + \max_{1 \le i \le n} [U_i + (b_k - \hat{b}_k) X_i] \end{aligned} \tag{3.9}$$

and

$$P(\max_{1 \le i \le n} [Y_i - \hat{b}_k X_i] - a_k \le z_n) = P(\max_{1 \le i \le n} [U_i + (b_k - \hat{b}_k) X_i] \le z_n) \ .$$

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 $P(|(b_k - \hat{b}_k) X_i| \le \epsilon_n) \to 1$ for a constant K. We take $z_n = \epsilon_n + \epsilon_n$ (or $z_n = z_n - \epsilon_n$) and apply the arguments of the last part of the proof of Theorem 1 to find

$$\hat{a}_k - a_k \xrightarrow{p} 0 \ . \tag{3.10}$$

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

By constructing the estimated efficiency frontier as

$$\hat{h}_m(x) = \hat{a}_k + \hat{b}_k x \ \ (\text{any } x \in I_k^{(m)}) \ , \tag{3.11}$$

we have the 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 have the following

result on $\hat{b}_k$ and $\hat{a}_k$.

Theorem 3 : Assume that U_i (≤ 0) is a sequence of i.i.d. negative-exponential random variables with λ (> 0) and X_i are bounded. In the present model, we consider the case when $n \longrightarrow \infty$ $(n(1), n(2) \to +\infty)$. Then we have the asymptotic distributions as follows,
(i) As $n \to \infty$

$$n(\hat{b}_k - b_k) \xrightarrow{w} Z_b = \lambda_1 Z_1 - \lambda_2 Z_2 \ , \tag{3.12}$$

where Z_i $(i = 1, 2)$ follows $G(\lambda) = e^{\lambda z_i}$ $(i = 1, 2)$. The distribution Z_b follows $G_b(z) = [\lambda_1/(\lambda_1 + \lambda_2) \exp[\frac{\lambda}{\lambda_1} z]$ $(z < 0)$, $G_b(z) = [-\lambda_2/(\lambda_1 + \lambda_2)[1 - \exp[\frac{\lambda}{\lambda_2} z]$ $(z \geq 0)$, where $\lambda_1 = [1/(X_M - X_L)][\lim_{n,n(1)\to\infty} \frac{n}{n(1)}]$ and $\lambda_2 = [1/(X_M - X_L)][\lim_{n,n(2)\to\infty} \frac{n}{n(2)}]$ $(\lambda_1 > 0, \lambda_2 > 0)$.
(ii) As $n \to \infty$

$$n(\hat{a}_k - a_k) \xrightarrow{w} Z_a \ , \tag{3.13}$$

where Z_a follows $G_a(z) = \exp[\lambda z_a]$ $(z_a \leq 0)$.

Proof of Theorem 3 : For the asymptotic distribution of $\hat{b}_k$, let $Z_{1n} = n(1)[\max_{I_1} Y_i - (a_k + b_k \bar{X}_L)]$ and $Z_{2n} = n(2)[\max_{I_2} Y_i - (a_k + b_k \bar{X}_M)]$. Then

$$n(\hat{b}_k - b_k) = \frac{n}{\bar{X}_M - \bar{X}_L}[\frac{Z_{1n}}{n(1)} - \frac{Z_{2n}}{n(2)}] = \lambda_{1n} Z_{1n} - \lambda_{2n} Z_{2n} \ , \tag{3.14}$$

where $\lambda_{1n} = \frac{n}{n(1)(\bar{X}_M - \bar{X}_L)}$ and $\lambda_{2n} = \frac{n}{n(2)(\bar{X}_M - \bar{X}_L)}$.
Because the asymtotic distribution of Z_{1n} and Z_{2n} is given by

$$G(z_1, z_2) = \exp[\lambda(z_1 + z_2)] \ (z_1 \leq 0, z_2 \leq 0) \ ,$$

where $\lambda_1 = \lim_{n\to\infty} \lambda_{1n}$ and $\lambda_2 = \lim_{n\to\infty} \lambda_{2n}$, We need some care on the asymptotic distribution of $\hat{b}_k$ because $Z_1 \leq 0$ and $Z_2 \leq 0$ and $Z = \lambda_1 Z_1 - \lambda_2 Z_2$ can take positive and negative values. When $Z = \lambda_1 Z_1 - \lambda_2 Z_2 \geq 0$, $\{Z \leq z\}$ and $Z_1 \leq 0$ imply $(\lambda_1 Z_1 - z)/\lambda_2 \leq Z_2 \leq (\lambda_1/\lambda_2) Z_1$. When $Z = \lambda_1 Z_1 - \lambda_2 Z_2 \leq 0$, $\{Z \leq z\}$ and $Z_2 \leq 0$ imply $Z_1 \leq (\lambda_2 Z_2 + z)/\lambda_1$. Hence we need to consider two cases, separately.
For $z < 0$ is given by

$$\begin{aligned} P(Z \leq z) &= P(Z_1 - \frac{\lambda_2}{\lambda_1} Z_2 \leq \frac{1}{\lambda_1} z) \\ &= \int_{-\infty}^{0} \left[\int_{-\infty}^{(\lambda_2 z_2 + z)/\lambda_1)} \lambda^2 \exp \lambda(z_1 + z_2) dz_1 \right] dz_2 \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \exp[\frac{\lambda}{\lambda_1} z] \ . \end{aligned}$$

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

$$P(Z \leq z) = \int_{-\infty}^{0} \left[\int_{(\lambda_1 z_1 - z)/\lambda_2}^{(\lambda_1/\lambda_2) z_1} \lambda^2 \exp \lambda(z_1 + z_2) dz_2 \right] dz_1$$

$$
\begin{aligned}
&= \int_{-\infty}^{0} \lambda \exp(\lambda z_1)[\exp(\lambda(\lambda_1/\lambda_2)z_1) - \exp(\lambda((\lambda_1 z_1 - z)/\lambda_2)z_1)]dz_1 \\
&= \frac{\lambda_2}{\lambda_1 + \lambda_2}[1 - \exp(-\frac{\lambda}{\lambda_1}z)] \ .
\end{aligned}
$$

To obtain the asymptotic distribution of $\hat{a}_k$, we prepare the next lemma.

Lemma 3.1 : Let $Z_{1n} = n(1)\max_{I_1} U_i$, $Z_{2n} = n(2)\max_{I_2} U_i$, and $Z_n = n\max_{I_1 \cup I_2} U_i$. Then the asymptotic distribution of (Z_{1n}, Z_{2n}, Z_n) is given by $G_\lambda(z_1, z_2, z)$.

Proof : We evaluate the joint probability and its approximation when $n(1), n(2) \longrightarrow \infty$ $(n \to \infty)$ as

$$
\begin{aligned}
& P(Z_{1n} \le z_1, Z_{2n} \le z_2, Z_n \le z) \\
= & \prod_{i=1}^{n(1)} P(U_i \le z_1/n(1), U_i \le z/n) \times \prod_{i=n(1)+1}^{n} P(U_i \le z_2/n(2), U_i \le z/n) \\
= & \prod_{i=1}^{n(1)} P(U_i \le \min(z_1, c_1 z)/n(1)) \times \prod_{i=n(1)+1}^{n} P(U_i \le \min(z_2, c_2 z)/n(2)) \\
\sim & \exp\{\lambda[\min(z_1, c_1 z) + \min(z_2, c_2 z)]\} \ ,
\end{aligned}
$$

where $c_1 = \lim_{n, n(1)\to\infty} n(1)/n$ (> 0) and $c_2 = \lim_{n, n(2)\to\infty} n(2)/n$ (> 0).
(**End of the proof of Lemma 3.1**)

For the asymptotic distribution of $\hat{a}_k$, we use the asymptotic distribution of $\hat{b}_k$, $\hat{a}_k = \max_{i=1,\cdots,n}[Y_i - \hat{b}_k X_i]$,

$$
\hat{a}_k - a_k = \max_{i=1,\cdots,n}[Y_i - a_k - b_k X_i + (b_k - \hat{b})X_i]
$$

and

$$
P(\hat{a}_k - a_k \le z_n) = P(\max[U_i + (b_k - \hat{b}_k)X_i] \le z_n) \ .
$$

We can use the fact that $n[\hat{b}_k - b_k] \xrightarrow{w} Z_b$ and X_i $(i = 1, \cdots, n)$ is bounded. Then we can take ϵ_n such that

$$
P(|[\hat{b}_k - b_k]X_i| \le \epsilon_n \text{ for any i}) \to 1
$$

and $\epsilon_n \to 0$ as $n \to \infty$.
Also

$$
\begin{aligned}
P(\max[U_i + \epsilon_n] \le z_n) &= \prod_{i=1}^{n} P(U_i \le z_n - \epsilon_n) \\
&= \exp[\lambda n(z_n - \epsilon_n)] \ .
\end{aligned}
$$

Then, by setting $(z_n - \epsilon_n) = z/n$, we have the result in Theorem 3.
(**Q**.**E**.**D**.)

It is important to notice that the order of convergence 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.

For the piece-wise linear function $h_m(x)$, we set $X = x$, and by using Lemma 3-1, we have the following asymptotic distribution.

Corollary 3.1 : Under the same setting with $x = \bar{X}$, $\hat{h}_m(x) - h_m(x) \xrightarrow{p} 0$, and the asymptotic distribution is given by

$$n[\hat{h}_m(x) - h_m(x)] \xrightarrow{w} Z_h = Z_a + (\lambda_1 Z_1 - \lambda_2 Z_2)x \ , \tag{3.15}$$

where the joint distribution of (Z_a, Z_1, Z_2) is given by

$$G_\lambda(z, z_1, z_2) = \exp\{\lambda[\min(z_1, \lambda_1 z) + \min(z_2, \lambda_2 z]\} \ \ (z_1 \leq 0, z_1 \leq 0, z \leq 0) \ . \tag{3.16}$$

The asymptotic distribution depends on the unknown parameter λ (> 0). It may be natural to use the residuals $\hat{U}_i = Y_i - \hat{a}_k - \hat{b}_k X_i$ 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

One important assumption on the 2nd method in Section 3 is that the inefficiency factor U follows the negative-exponential distribution. In this section we assume that the inefficiency is a sequence of i.i.d. 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

$$Y_{kj} = b_k X_k + U_{ij} \ (k = 1, \cdots, m; j = 1, \cdots, n_k) \tag{4.1}$$

where U_{kj} (≤ 0) is a sequence of i.i.d. random variables with the distribution function F and we have 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 $f(0)$ is bounded. We use

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

as $n_k \to \infty$ if we take $z_n = z/n_k + b_k X_k$ and $\bar{F}(x) = 1 - F(x)$.
Then, by using the Taylor expansion of $\bar{F}(x)$ around $x = 0$ ($\bar{F}(0) = 0$), as $n_k \to \infty$ $(k = 1, \cdots, m)$

$$P(n_k[\max_{j=1,\cdots,n_k} Y_{kj} - b_k X_k] \leq z) \longrightarrow \exp[f(0)z] \ \ (z \leq 0) \ , \tag{4.2}$$

provided that $f(0)$ is bounded.
When $F(u)$ 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 a general situation when $f(x)$ diverges at $x = 0$. A typical case may be the pareto-type distribution, and then we need to assume that

$$\bar{F}(-x^{-1}) = x^{-\alpha} L(x) \ , \tag{4.3}$$

where $L(x)$ is a slowly varying function and $\alpha > 0$. This formulation is the standard assumption in the statistical extreme value theory (SEVT) in statistics.
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 $n_i \to \infty$

$$P(c(n_k)[\max_{j=1,\cdots,n_k} Y_{kj} - b_k X_k] \leq z) \longrightarrow \exp[-(-z)^{\alpha}] \ \ (z \leq 0) \ , \tag{4.4}$$

where $\alpha > 0$.
This asymptotic distribution has been known as the Weibull-type 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.

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 $I_{k_j}^{(m_j)} = (w_{j1}^{(k_j)}, w_{j2}^{(k_j)}]$ ($w_{j1}^{(k_j)} \leq w_{j2}^{(k_j)}$), $0 \leq w_{j1}^{(0)} < w_{j1}^{(1)} < \cdots < w_{j1}^{(m_j)}$ and $0 \leq w_{j2}^{(0)} < w_{j2}^{(1)} < \cdots < w_{j2}^{(m_j)}$. For $k = (k_1, k_2)'$, $m = (m_1, m_2)'$ and $I_{k_j}^{m_j} = (w_{j1}^{k_j}, w_{j2}^{k_j}]$ $(j = 1, 2)$, we set the $(k_1, k_2)-$th region by $\mathbf{I}_k^{(m)} = I_{k_1}^{(m_1)} \times I_{k_2}^{(m_2)}$. We estimate the hyperplanes of the form

$$h_m(\mathbf{X}) = a_k + b_{1k} X_1 + b_{2k} X_2 \tag{5.1}$$

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

Let vectors $\mathbf{x} = (x_1, x_2)'$, $\mathbf{x}(1) = (x_1(1), x_2(2))'$ and $\mathbf{x}(2) = (x_1(2), x_2(2))'$ be in $\bigcup_k \mathbf{I}_k^{(m)}$ and let non-negative scalars λ_j $(j = 1, 2)$. Then the concavity restrictions imply that

$$h_m(\mathbf{x}) \geq \lambda_1 h_m(\mathbf{x}(1)) + \lambda_2 h_m(\mathbf{x}(2)) \tag{5.2}$$

for any $\mathbf{x} = \lambda_1 \mathbf{x}(1) + \lambda_2 \mathbf{x}(2)$ and $\lambda_1 + \lambda_2 = 1$.
It is straightforward to check these conditions numerically at every estimation, but there may be some complications in their numerical evaluations. Then we have used 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 either $I_1(1)$ or $I_2(1)$ and take two intervals $I_1(1) = I_1(2) \cup I_2(2)$ or $I_2(1) = I_1(2) \cup I_2(2)$. 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 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 regression 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. (We have taken m such that m_1 and m_2 are less than $0.1\times$ (sample size).)

The first estimation method of a and b

We can 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 such that $n(p, q) = n_{k_1,k_2}(p, q)$ $(p, q = 1, 2)$. 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 estimation 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

$$\hat{\mathbf{b}}_k^{LS} = [\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^{`}]^{-1} [\sum_{i=1}^{n} \mathbf{X}_i Y_i] \ , \tag{5.3}$$

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

$$\hat{a}_k^{LS} = \min_{i=1,\cdots,n} \{a | a + \hat{b}_k^{LS} X_i \geq Y_i\} \ . \tag{5.4}$$

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.

The second estimation method of a and b

We can apply the 2nd estimation method based on the SEVT method explained in Section 3 with the concavity restrictions. As for an illustration and we divide the regions $I_k^{(m)}(1,1) = (w_{11}^{(k_1)}, x_1] \times (w_{21}^{(k_2)}, x_2]$, $I_k^{(m)}(1,2) = (x_1, w_{11}^{(k_1)}] \times (w_{21}^{(k_2)}, x_2]$, $I_k^{(m)}(2,1) = (w_{11}^{(k_1)}, x_1] \times (x_2, w_{21}^{(k_2)}]$, and $I_k^{(m)}(2,2) = (x_1, w_{11}^{(k_1)}] \times (x_2, w_{22}^{(k_2)}]$.

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

$$\begin{aligned} P(\max_{I(2,1)} Y_i \le z_n) &= P(\max_{I(2,1)} [U_i + a + b_1 X_{1i} + b_2 X_{2i}] \le z_n) \\ &= \prod_{i=1}^{n(2,1)} P(U_i + a + b_1 X_{1i} + b_2 X_{2i} \le z_n) \\ &= \exp\{\lambda n(2,1) z_n - n(2,1) a - b_1 \sum_{i=1}^{n(2,1)} X_{1i} - b_2 \sum_{i=1}^{n(2,1)} X_{2i}) \\ &= \exp\{\lambda n(2,1)[z_n - a - b_1 X_1(2,1) - b_2 X_2(2,1)]\} \ . \end{aligned}$$

Then, by using the same arguments in Section 3, we have

$$\max_{I(2,1)} Y_i - [a + b_1 X_1(2,1) + b_2 X_2(2,1)] \xrightarrow{p} 0 \ .$$

Similarly, we find that $\max_{I(1,1)} Y_i - [a + b_1 X_1(1,1) + b_2 X_2(1,1)] \xrightarrow{p} 0$ and $\max_{I(1,2)} Y_i - [a + b_1 X_1(1,2) + b_2 X_2(1,2)] \xrightarrow{p} 0$.
By using the above relations,

$$[Y_M(2,1) - Y_M(1,1)] - b_1[X_1(2,1) - X_1(1,1)] - b_2[X_2(2,1) - X_2(1,1)] \xrightarrow{p} 0$$

and

$$[Y_M(1,2) - Y_M(1,1)] - b_1[X_1(1,2) - X_1(1,1)] - b_2[X_2(1,2) - X_2(1,1)] \xrightarrow{p} 0 \ .$$

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

$$\begin{bmatrix} Y_M(2,1) - Y_M(1,1) \\ Y_M(1,2) - Y_M(1,1) \end{bmatrix} = \begin{bmatrix} X_1(2,1) - X_1(1,1) & X_2(2,1) - X_2(1,1) \\ X_1(1,2) - X_1(1,1) & X_2(1,2) - X_2(1,1) \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \ .$$

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

$$\hat{a} = \min_{i \in I_{k_1,k_2}} \{a | a + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} \ge Y_i\} \ . \tag{5.5}$$

We have found that

$$\left[\begin{array}{c} Y_M(2,1) - Y_M(1,1) \\ Y_M(1,2) - Y_M(1,1) \end{array} \right] - \left[\begin{array}{cc} X_1(2,1) - X_1(1,1) & X_2(2,1) - X_2(1,1) \\ X_1(1,2) - X_1(1,1) & X_2(1,2) - X_2(1,1) \end{array} \right] \left[\begin{array}{c} \hat{b}_1 - b_1 \\ \hat{b}_2 - b_2 \end{array} \right]$$
$$\xrightarrow{p} \quad 0 \ .$$

For instance, if we assume that

$$\text{rank} \left[\begin{array}{ccc} 1 & X_1(2,1) & X_2(2,1) \\ 1 & X_1(1,2) & X_2(1,2) \\ 1 & X_1(1,1) & X_2(1,1) \end{array} \right] = 3 \ , \tag{5.6}$$

then $\hat{b}_1 - b_1 \xrightarrow{p} 0$ and $\hat{b}_2 - b_2 \xrightarrow{p} 0$.
Let $Z_n(2,1) = n(2,1)[\max_{I(2,1)} Y_i - (a + b_1 X_1(2,1) + b_2 X_2(2,1))]$,
$Z_n(1,2) = n(1,2)[\max_{I(1,2)} Y_i - (a + b_1 X_1(1,2) + b_2 X_2(1,2))]$,
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, we have the limiting exponential random variables $Z(2,1)$, $Z_(1,2)$, and $Z(1,1)$ with the joint distribution

$$G(z_{21}, z_{12}, z_{11}) = \exp[\lambda(z_{21} + z_{12} + z_{11})] \ \ (z_{21} \leq 0, z_{12} \leq 0, z_{11} \leq 0) \ .$$

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

$$\begin{aligned} \mathbf{Z}_b \ = \ & \left[\begin{array}{cc} X_1(2,1) - X_1(1,1) & X_2(2,1) - X_2(1,1) \\ X_1(1,2) - X_1(1,1) & X_2(1,2) - X_2(1,1) \end{array} \right]^{-1} \\ & \left[\begin{array}{ccc} \lambda(2,1) & 0 & -\lambda(1,1) \\ 0 & \lambda(1,2) & -\lambda(1.1) \end{array} \right] \left[\begin{array}{c} Z(2,1) \\ Z(1,2) \\ Z(1,1) \end{array} \right] \ , \end{aligned}$$

where $\lambda(2,1) = \lim_{n\to\infty} n/n(2,1)$, $\lambda(1,2) = \lim_{n\to\infty} n/n(1,2)$, and
$\lambda(1,2) = \lim_{n\to\infty} n/n(1,1)$ as $n, n(2,1), n(1,2), n(1,1) \to \infty$.
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

$$n[\hat{h}_m(\mathbf{x}) - h_m(\mathbf{x})] \xrightarrow{p} Z_h = Z_a + \mathbf{Z}_b^{`}\mathbf{x} \ , \tag{5.7}$$

where $\mathbf{x} = (x_1, x_2)^{`}$.
This expression can be directly generalized 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 multidimension cells. To illustrate an alternative method, we use

the case when $p = 2$ and we denote each cell as $I(j,k)$ $(j,k = 1,2)$. We also use the notations such that for $j,k = 1,2$ $\bigcup_k I(j,k) = I(j,\cdot)$ and $\bigcup_j I(j,k) = I(\cdot,k)$.
To cope with this problem, we use the relation

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

Then we can develop the similar evaluation except the fact that the resulting limit random variables $Z_n(j,\cdots)$ and $Z_n(\cdot,k)$ $(j,k = 1,2)$ 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

$$G(z_{j,\cdot}, z_{\cdot,k}) = \exp\{\lambda \sum_{j,k} [z_{j,\cdot} \wedge z_{\cdot,k}]\} \ (j,k = 1,2), \tag{5.8}$$

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

$$\begin{aligned} \mathbf{Z}_b^* \ = \ & \begin{bmatrix} X_1(2,\cdot) - X_1(1,\cdot) & X_2(2,\cdot) - X_2(1,\cdot) \\ X_1(\cdot,2) - X_1(\cdot,1) & X_2(\cdot,2) - X_2(\cdot,1) \end{bmatrix}^{-1} \\ & \times \begin{bmatrix} \lambda(2,\cdot) & -\lambda(1,\cdot) & 0 & 0 \\ 0 & 0 & \lambda(\cdot,2) & -\lambda(\cdot,1) \end{bmatrix} \begin{bmatrix} Z(2,\cdot) \\ Z(1,\cdot) \\ Z(\cdot,2) \\ Z(\cdot,1) \end{bmatrix}. \end{aligned} \tag{5.9}$$

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.

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

$$Y_i = a_k + \sum_{j=1}^{p} b_{jk} X_{ji} + U_i \ \ (i = 1,\cdots,n) \ , \tag{5.10}$$

where $U_i \ \leq 0$.

6. Efficient Frontier and Measurement Errors

There are cases when we should not ignore the measurement errors in inputs and outputs. Let V_i be the measurement errors for the i-th observation. When $V_i < 0$, it may not be possible to distinguish it from the inefficiency term, which does not take any positive value. Hence it is reasonable to 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

$$f(X_i) = f(X_i)^*(1+c) \ , \tag{6.1}$$

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, their roles could not be ignorable.

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, and then we 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. It is appropriate to treat Kanpo-Seimei as an outlier and should be deleted, which is not discussed in detail here. Here we just 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

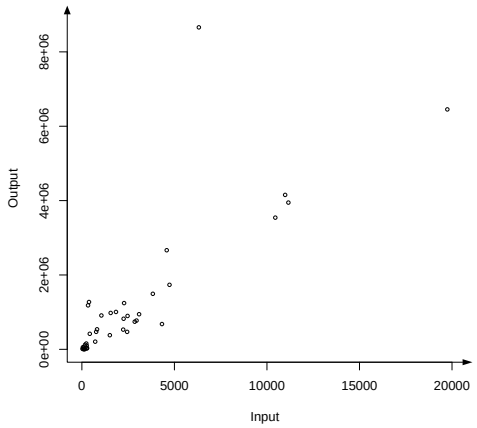

Figure 3: An outlier situation : In the life-insurance industry in Japan, there is an outlier and there are some reasons why it is.

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 have developed a SDEA method based on the statistical modeling of linear regression and extreme value distributions, which may be new to both the operations reserach 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, we have used the linear regression based method for estimating coefficients. When the number of data is large, however, we have shown that we have some efficiency gain in the statistical estimation if we use the statistical extreme value (SEVT) method.

There are a number of problems in the SDEA method remained to be investigated. The statistical models treated in this paper can be generalized to several directions including multivariate inputs and outputs. Then, 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 have a huge number of data, it may be possible to use may explanatory variables.

Another important statistical issue would be that there can be several procedures

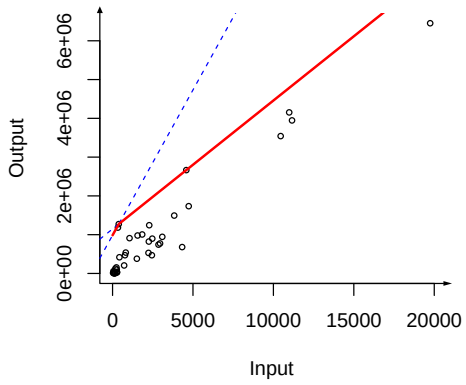

Figure 4: An estimeted frontier : Input is Workers and output is Income.

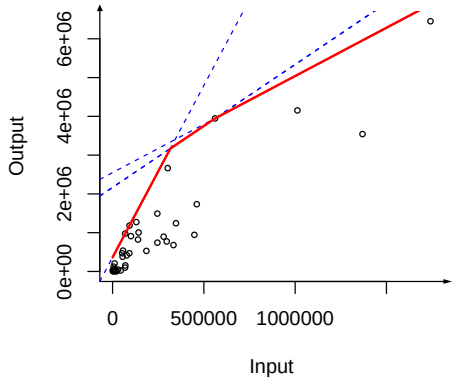

Figure 5: An estimeted frontier : Input is Capital and output is Income.

to choose the number of intervals (m) in data analysis and we need to develop some criterion of selecting the number of interval nodes (m) in an optimal way given a finite number of data.

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 analysis: 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).