SSE-DP-2026-3

統計的データ包絡分析
－ SDEA －

国友直人・趙宇

統計数理研究所・東京理科大学

2026 年 3 月

SSE-DP(ディスカッションペーパー・シリーズ)は以下のサイトから無料で入手可能です。
https://stat-expert.ism.ac.jp/training/discussionpaper/

SSE-DP-2026-3

Statistical Data Envelopment Analysis

by

Naoto Kunitomo and Yu Zhao

The Institute of Statistical Mathematics・Tokyo Science University

March 2026

(Summary)

This is a revised version of SSE-DP-2022-4 written in English. In operations research and management sciences, Data Envelopment Analysis (DEA) has been known as one of important tools. We develop Statistical Data Envelopment Analysis (SDEA), which seems to be new to operations research literature 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. Contrary to some related studies, the hidden function is not necessarily differentiable. The purpose of data analysis is to estimate the un-known function, called the efficiency frontier, nonparametrically based on the set of observed data of inputs and outputs. The key idea is to use the non-parametric 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.

統計的データ包絡分析（SDEA）*

国友直人 †

趙宇 ‡

2026 年 3 月 5 日

概 要

オペレーションズ・リサーチ（Operations Research）および経営科学 (マネジメント・サイエンス) の分野では，データ包絡分析（Data Envelopment Analysis; DEA）は重要な分析法として広く認識されている．本研究では，オペレーションズ・リサーチおよび統計学の分野における試みとして，統計的データ包絡分析（Statistical Data Envelopment Analysis; SDEA）を提案する．本稿では観測データが入力の増加的かつ (必ずしも微分可能とは限らない) 凹関数と非正の確率的ノイズの和として表される基本的な統計的 DEA モデルを考察する．ノイズ項は入力と出力の関係における非効率性として解釈される．データ分析の目的は，観測された入力および出力データに基づき，未知の関数（フロンティア関数と呼ぶ）をノンパラメトリックに推定することである．本研究では線形回帰および統計的極値論（Statistical Extreme Value Theory; SEVT）の統計的手法を利用する．応用例として日本の生命保険業に関するデータ分析を例示する。

鍵言葉

SDEA (Statistical Data Envelopment Analysis）, 非効率性, 微分可能とは限らない凹関数, 回帰 DEA, 極値 DEA, 日本の生命保険業

---

*Version-J-2026-3-5. この論文は Discussion Paper SSE-DP-2022-4.pdf(英文) の改訂稿である. (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.) 原論文に対する湯浅良太氏, 吉田靖氏, 椿広計氏からのコメントに感謝する。この研究は「統計エキスパート人材育成事業」(Consortium for training experts in statistical sciences at the ISM) の一環として行われたが, JSPS Grant22K01428 の支援も受けた.

†統計数理研究所

‡東京理科大学経営学部

1　はじめに

オペレーションズ・リサーチ（Operations Research; OR）およびマネジメント・サイエンスの分野では、データ包絡分析（Data Envelopment Analysis; DEA）は重要な分析手法として広く認識されている。オペレーションズ・リサーチにおけるDEAは、しばしば数理計画法の応用として議論されているが、既存の方法や発展については、例えばCooper, Seiford, and Tone (2007) を参照されたい。これに対し、経済経済学の分野では、Aigner, Lovell, and Schmidt (1977) 以降、生産フロンティアのパラメトリック推定の問題として検討されている。これは、企業の行動を利潤最大化や費用最小化として理解するミクロ経済学の基本的枠組みに関連しており、計量経済学においては基礎的な問題と見なせる。ミクロ経済学におけるフロンティア関数の凹関数と準凹関数の役割の説明については例えば Mas-Colell, Whinston and Green (1995) を参照されたい。Aigner et al. (1977) では半正規分布などの切断分布をもつ誤差項を含むパラメトリックな線形回帰モデルの利用を提案、最尤推定（Maximum Likelihood Estimation; MLE）により効率的フロンティア関数としての生産関数を推定している。こうしたフロンティア関数の推定に関する計量経済学的研究の詳細は、例えばGreene (2003) の第17章に詳しい。なおオペレーションズ・リサーチと計量経済学で発展した分析法の目的は類似しているものの、問題解決のために用いる理論的枠組みおよび数理的手法には大きな違いがある。既存の文献では一方のOR分野では確率分布にしたがうノイズをあまり考えない線形・非線形計画問題の応用としての方法が展開されているが、他方、計量経済学における生産関数の推定問題ではパラメトリック線形・非線形回帰問題の応用として扱われていることが多い。

この論文では、オペレーションズ・リサーチ、計量経済学、および統計科学における新しい試みとして、統計的データ包絡分析（Statistical Data Envelopment Analysis; SDEA）を提案する。まず、観測データが増加的かつ（必ずしも微分可能とは限らない）凹関数と非負値をとる確率的ノイズの和として表される基本的な統計的DEAモデルを考察する。そしてこの確率的ノイズ項を入力と出力の関係における非効率性 (inefficiency) を意味する。統計的データ分析の目的は、観測データ集合に基づいて未知の関数 (包絡関数, フロンティア関数と呼ばれている)、すなわち効率的フロンティア（efficient frontier）をノンパラメトリックに推定することである。本研究では、未知のフロンティア関数 (包絡関数) を推定するために、線形回帰分析および統計的極値理論（Statistical Extreme Value Theory; SEVT）に基づく統計的手法を考察する。第一の推定法として、線形回帰に基づく回帰DEA法を導入する。この方法は単純で実装も容易であるが、大標本の場合には推定効率は十分高くなく、改善の余地があることが分かる。次に、第二の推定法として、SEVTに基づく極値DEA法を導入する。SEVTは統計学においてよく知られているとまでは云えないが、この方法を利用するとデータ数が多い場合という設定の下では第一の方法よりも推定

量の収束速度が改善できることが分かる。ただし、サンプルサイズが多くない場合にはSEVT に基づく推定法が必ずしも満足できる結果を与えないことありうることを数値例の議論を通じて指摘する。ここで提案する回帰 EDA 法と極値 EDA 法では未知であるフロンティア関数について単調かつ凹関数と云う性質の他には特に微分可能性の条件などは仮定しないが、回帰 EDA では誤差項の分散の存在は仮定する。極値 EDA では期待値や分散の存在などは仮定しなくてよい. 本稿ではゼロ点における密度の有界性と領域についての制約を課すが、3 節で述べるように条件を緩めることは可能である。また回帰 EDA 法の方が容易に実用に供されるが,3 節で示すようにデータ数が多い場合には極値 EDA 法の方が推定効率が高まる可能性がある。

本稿で議論する統計モデルに関連した研究は近年では OR や経済学などの分野以外でも散見される。例えば Millimet and Parmeter (2021) は政治・社会学 (Political Analysis) の分野では災害や犯罪などの公表データにはバイアスが存在する可能性が高いとして one-sided measurement error の分析の必要性を議論している。ただし未知のフロンティア関数の形状については生産関数や DEA のような理論的考察を行っていない。また Jirak, Meister, and Reiss(2014) は one-sided errors が存在する場合のノンパラメトリック統計分析の方法を提案、推定量の効率限界を検討している。本論文の問題設定とは異なり、関連する問題として境界関数が 1 次元の説明変数の滑らかな関数に Lipschitz タイプの条件、微分可能性を仮定した上で一般的な統計的モデルにおける誤差分布の設定と推定の効率境界などについて議論している。ここでは彼らが得た結果は未知関数の滑らかさという形状について分析上で置く数学的仮定に結果が強く依存する、ことを指摘しておくことにとどめる。

この論文の主な目的は、DEA の問題を統計科学における統計的問題として整理し、未知関数の形状については（微分可能性を仮定せず）凹関数という仮定のみを利用するとともに、比較的容易に実現できる統計的推定法を提案することである。さらに提案する方法を SDEA(Sstatistical EDA) と呼び、その統計理論の側面, およびシミュレーション的結果を報告する。特にノンパラメトリックな SEDA 法として回帰 EDA 法と極値 EDA 法を導入し、また応用例として日本の生命保険業界のデータ分析を例示する。なおこの応用ではデータ数が約 40 と DEA の分析としては比較的小さいため、回帰ベースの推定法を適用した。本研究のアプローチは従来のオペレーションズ・リサーチおよびマネジメント・サイエンスの枠組みとはかなり異なることを踏まえ、また統計科学のコミュニティーを主な対象として EDA の問題を理解するため、まず説明変数が 1 次元という基本的な場合を比較的に詳しく説明し、一般的な場合はより簡潔に説明する。また統計的極値論についての知識を仮定せず、極値分布や極値 EDA で得られる極限分布についても言及する。

本論文の構成は以下の通りである。第 2 節では SDEA の定式化と単純な場合における回帰 DEA を説明する。第 3 節ではデータ数が多い場合に適用可能な極値 DAE 導入し、繰り返し測定（repeated measurement）の場合における SDEA モデルと統

計的極値論との関係を議論する。第 4 節では複数の説明変数を含む一般的な SDEA モデルに対して推定法を説明する。第 5 節ではシミュレーションの結果を報告し、第 6 節では日本の生命保を題材に効率性について回帰 DEA による分析結果を例示し、最後に第 7 節でまとめを述べる。数理的補論に本稿で述べた数理的結果の証明を与えておく。

2 SDEA 法

2.1 データ包絡分析の統計的問題

本研究では，問題を統計的 DEA（Data Envelopment Analysis）モデルのノンパラメトリック推定として定式化する。出力および入力をそれぞれ Y および X とし，これらはいずれも非負の値をとるものとする。ここではフロンティア関数 $h(\cdot)$ について連続性は仮定するが、Lipschitz タイプの条件や微分可能性は特に仮定しない。この点は SDEA を他のノンパラメトリックな統計的問題と区別する意味で重要であるので特に強調しておく。むろんフロンティア関数が滑らかで二階微分可能なら $h' > 0$ および $h'' < 0$ を意味する。未知のフロンティア関数からの非効率性項を表す確率変数 U を導入し、$h(X)$ が与えられたもとで出力 Y は二項の和として表されるとすると,

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

標準的な DEA では，X と Y は任意の実数値をとり得るが，実際の応用では，X および Y の有限個の観測値しか得られない。標本サイズを N とする。観測される出力・入力をそれぞれ $Y_i\ (i = 1, \cdots, N)$ および $X_i\ (i = 1, \cdots, N)$ とし，これらはいずれも非負値をとる。ここでは入力水準 X に対して，区間 $\mathrm{I}_k^{(n)}$ 上の増加かつ凹の区分線形フロンティア関数 $h_m(X)$ を次のように設定しよう。

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

ここで，$\mathrm{I}_k^{(n)} = (w_1^{(k)}, w_2^{(k)}]\ (w_1^{(k)} < w_2^{(k)})$ とし，$0 \leq w_1^{(1)} < \cdots < w_1^{(m)}$ および $0 \leq w_2^{(1)} < \cdots < w_2^{(m)}$ が成り立つものとするが、本稿では X_i が有界な変数である場合に限定して議論を進める。
$h(x)$ の凹性を反映して，係数には次の単調性制約を課す。

$$0 \leq a(1) \leq \cdots \leq a(m)\ , \quad b(1) \geq \cdots \geq b(m) \geq 0\ . \tag{2.3}$$

また，$U_i\ (i = 1, \cdots, N)$ は X が与えられたもとでの独立同分布（i.i.d.）にしたがう非正確率変数列であると仮定する。典型的な場合としては U_i がパラメータ $\lambda > 0$

をもつ負の指数分布にしたがうとすると，$F(u) = \mathrm{P}(U_i \leq u) = \exp[\lambda u]$ $(u \leq 0)$. である。このとき統計モデルは次のように表される。

$$Y_i = h_m(X_i) + U_i \ \ (i = 1, \cdots, N) \ . \tag{2.4}$$

ここで $h_m(X_i)$ が X_i について単調非減少な区分線形凹関数であり，U_i が非正の実数値のみをとるという制約があるが、推定の対象であるフロンティア関数 $h(X)$ は研究者には未知である。有限個のデータセット (X_i, Y_i) $(i = 1, \cdots, N)$ が与えられた場合，未知関数 $h_m(x)$ を推定できるのは $m = m_N < N$ のときに限られる。区間 $\mathrm{I}_k^{(n)}$ $(k = 1, \cdots, m)$ は $\bigcup_{k=1}^m \mathrm{I}_k^{(n)} = (w_1^{(1)}, w_2^{(m)}]$ を満たすように選び，$\mathrm{I}_k^{(n)} = (w_1^{(k)}, w_2^{(k)}]$ 内のデータ数を n_k として

$$\sum_{k=1}^{m} n_k \geq N \tag{2.5}$$

が成り立つものとする。ここでは $\sum_{k=1}^m n_k > N$ の場合も許容するが、区間は重複して取りうることを意味している。ここでは入力変数 X が有界（またはいくつかの入力変数が有界）であり，区間の境界が事前に既知である場合を扱う。また，X の区間をランダムに選ぶことも可能であり、有限データに対してランダム区間を用いるより効率的な推定法が開発できる可能性があるが, 本研究では議論しないことにする。

各区間でのデータ数 n_k $(k = 1, \cdots, m)$ および m がともに大きい場合には，推定に関する漸近的性質を利用することができる。以下では，

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

であれば

$$\sup_x \left|\hat{h}_m(x) - h(x)\right| \xrightarrow{p} 0 \quad (m \to +\infty, \ n_k \to +\infty) \ . \tag{2.6}$$

となるので、区分線形関数 $\hat{h}_m$ の推定法を検討する。

ここでデータ数 N に対して滑らかな関数 $h(x)$ を推定する一つの実際的な方法としては，有限個の m 個の節点をとり、推定値 $\hat{h}_m(x)$ に基づき，スプライン関数などを用いることが考えられるが, これがOR分野での伝統的DEA法として解釈できる。ここで図1にこの問題の典型的な状況を示しておこう。例えば100社の企業が共通の技術 $Y = X^{0.3}$ $(X > 0)$ のもとで，入力 X と出力 Y を生産していると想定してみよう。産業あるいは市場の中には効率的な企業も存在するが，ほとんどの企業は非効率であり，非効率性を確率変数 U $(U \leq 0)$ で表そう。ここで U は非正値をとる連続確率変数であり、図1では U の乱数は負値をとる指数分布（$\lambda = 1/10.0$）から生成した。フロンティア関数 $h(X) = X^{0.3}$ の正確な形は未知であり，$Y(= h(X) + U)$ および h が非負かつ凹であることのみが知られているのが現実的だろう。このとき,

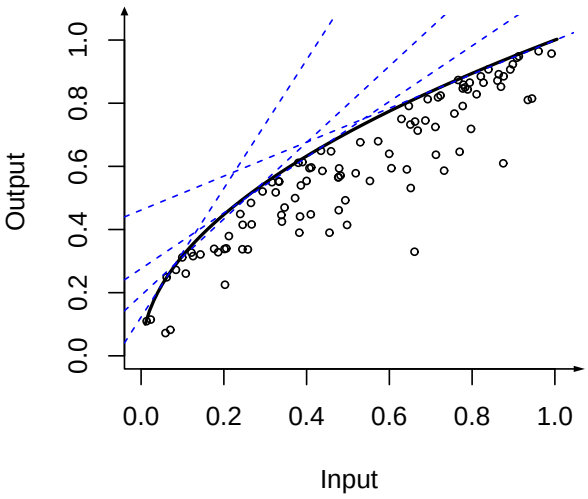

図 1: 典型的な SDEA の状況：シミュレーションによるデータ 100 個に基づく推定問題（$f(X) = X^{0.3}$ $(X > 0)$, $-U$ は数分布（$\lambda = 1/10.0$）により生成。）

データ集合 (X_i, Y_i) $(i = 1, \cdots, 100)$ から未知関数 h をノンパラメトリックに推定することが統計的な問題である。この際，ある X の近傍のデータ点群を用いて，真のフロンティア関数に接する複数の線分を局所的に描くことができる。図 1 には仮想的に推定した 4 本の接線を例示しておいた。

2.2 ノンパラメトリック推定法

本研究では二つのノンパラメトリック推定法を提案する。全体のデータから m 区間を選択, 第 k 区間 $\mathrm{I}_k^{(n)}$ において，データ数を $n = n_k$ $(k = 1, \cdots, m)$ とする (n は十分大きくとる)。ここで $\mathrm{I}_k^{(n)}$ 内での $h(X)$ の接線を推定する問題を考えるが, 任意の $X = x(> 0)$ に対して，

$$Y_i = a(k) + b(k)X_i + U_i \quad (i = 1, \cdots, n) \tag{2.7}$$

と表されるものとする。なお, 本稿では表記上の理由より各区間 k について混乱がなければ簡化して $a = a(k)$, $b = b(k)$，$a(k) + b(k)x \geq h(x)$，$X_i \in \mathrm{I}_k^{(m)} = (w_1^{(k)}, w_2^{(k)}]$, $a(k)$ および $b(k)$ を知パラメータとする。ここで未知関数 $h(x)$ を各 x における値を推定するために，接線関数 $a + bx$ を用いることを提案する。

2.3 回帰 DEA 推定

データサイズがそれほど大きくなくとも利用できる方法として, 各区間ごとに線形回帰モデルに基づく簡単で自然な推定法を説明しよう。第 k 区間において, 回帰式の傾き係数を

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

切片係数は次の最小値として定義する。

$$\hat{a}(k)^{LS} = \min_{i=1,\cdots,n} \{a \mid a + \hat{b}(k)X_i \geq Y_i\} , \tag{2.9}$$

なお $n = n_k$, $\bar{Y} = (1/n)\sum_{i=1}^{n} Y_i$, および $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ である。
ここでフロンティア関数には凹性が仮定されているので, 推定する係数に単調性制約を課す必要があり，次の条件を仮定する。

$$0 \leq \hat{a}(1)^{LS} \leq \cdots \leq \hat{a}(m)^{LS} , \quad \hat{b}(1)^{LS} \geq \cdots \geq \hat{b}(m)^{LS} \geq 0 .$$

推定された係数が区間内でこれらの制約を満たさない場合には，その区間の推定値を採用せず、区間の取り方を変更することが考えられる。次に誤差項に関するモーメント条件のもとで，漸近的な結果を示すが, 導出は数理的補論に与えておく。

定理 1. U_i (≤ 0) が i.i.d. な確率変数列であり，分散 $\mathbf{V}[U_i] = \sigma_u^2 < +\infty$，密度関数 $f(u)$ は $u = 0$ で有界かつ滑らかであり，X_i は有界であることを仮定する。
(i) 各区間 $\mathrm{I}_k^{(n)}$ において，$n\ (= n_k) \to \infty$ のとき，

$$\begin{bmatrix} \hat{a}(k)^{LS} - a(k) \\ \hat{b}(k)^{LS} - b(k) \end{bmatrix} \xrightarrow{p} \mathbf{0} . \tag{2.10}$$

(ii) さらに，$n\ (= n_k) \to \infty$ のとき，

$$\sqrt{n}[\hat{b}(k)^{LS} - b(k)] \xrightarrow{w} N(0, \sigma_b^2) , \tag{2.11}$$

ただし，$\sigma_b^2 = \frac{\sigma_u^2}{M_x}$，$M_x = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ が正定数となることを仮定する。
(iii) 区間を $\mathbf{I}_k^{(n)} = (X_L^{(n)}, X_M^{(n)}]$，$\lim_{n\to\infty} X_L^{(n)} = X_L$，$\lim_{n\to\infty} X_M^{(n)} = X_M$ と仮定する。確率変数 a^* を $\sqrt{n}[\hat{a}(k)^{LS} - a(k)]$ の極限分布とする。このとき，$a_L^* \leq a^* \leq a_M^*$，

$$a_M^* = -b_+^* X_L - b_-^* X_M , \quad a_L^* = -b_+^* X_M - b_-^* X_L$$

ととれる。ここで $b_+^* = b^*$ (if $b^* \geq 0$)，$b_-^* = b^*$ (if $b^* < 0$)，b^* は $\sqrt{n}(\hat{b}_k LS - b_k)$ の極限を表す確率変数であり, a_M^* の密度関数は次のように与えられる。

$$g_M(z) = \frac{1}{\sigma_b X_M}\, n\Big(-\frac{z}{\sigma_b X_M}\Big) \quad \text{if } z \geq 0, \tag{2.12}$$

$$g_M(z) = \frac{1}{\sigma_b X_L}\, n\Big(-\frac{z}{\sigma_b X_L}\Big) \quad \text{if } z < 0, \tag{2.13}$$

ここで $n(z)$ は標準正規分布の密度関数である。X_M と X_L を入れ替えることで，a_L^* の密度関数は $g_L(z) = \frac{1}{\sigma_b X_L} n(-z/(\sigma_b X_L))$ (if $z \geq 0$), $g_L(z) = \frac{1}{\sigma_b X_M} n(-z/(\sigma_b X_M))$ (if $z < 0$) となる。

なお a_M^* および a_L^* の極限分布をここでは区分正規分布（piecewise normal）と呼んでおく。
ここで関数 $h_m(x)$ は各 k について $a_L^* + b^* x = b_+^*(x - X_M) + b_-^*(x - X_L)$ および $a_M^* + b^* x = b_+^*(x - X_L) + b_-^*(x - X_M)$ により近似できる。
区間 $\mathbf{I}_k^{(n)}$ の幅が大きくなければ $(X_L + X_M)/2$ $(= X_k^c)$ を利用して, a^* の分布を $N(0, \sigma_b^2 X_k^{c2})$ で近似することができる。$\hat{b}(k) - b(k)$ および $\hat{a}(k) - a(k)$ の収束オーダーは $\sqrt{n}$ である。3 節で示すように，ある種の状況下では，誤差項に関するモーメント条件を課さずに，傾き $\hat{b}(k)$ および 切片 $\hat{a}(k)$ の推定において収束の次数を改善することが可能である。第 2 の推定法である極値 DEA 法では，追加条件のもとでは収束速度が n となり得ることが示される。

2.4 数値例

回帰 DEA 法を用いた SDEA の数値例を示しておく。図 2 は図 1 で利用したシミュレーションデータに基づいて回帰 DEA を利用して推定した区分線形関数 (フロンティア関数) を示している。この例では真のフロンティア関数は連続かつ凹であるが，観測データ数が有限であり負のノイズを含むため，いくつかの区間ではデータ上では非凹として観察される場合もある。回帰 DEA 法は，区分線形のフロンティア関数を用いることでかなり実用的に機能することが分る。この例では重複しない区間を用いて $m = 4$ としている。観測データが大きくない場合でも，この単純なケースが示すように，回帰 DEA 法は実用上では妥当な推定結果をもたらすことが多いことが分かる。

2.5 m の選択問題

実務上では推定に利用する区間数としての m を選ぶことは重要である。真のフロンティア関数が非線形の凹関数である場合，それを推定するには大きな m が必要になる。しかし有限標本では，推定係数が（2.3）に示した単調性制約を満たすという条件の下で，有限個の区間を設定する必要がある。そこで与えられたデータ集合に対して最適な m を見つけるために，次の手続きを提案する。

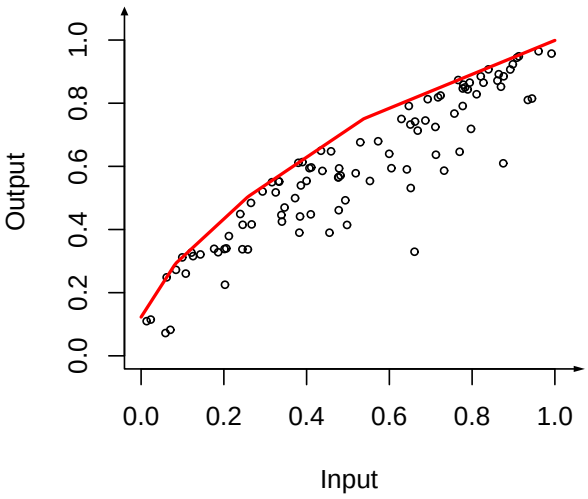

図 2: 回帰 DEA 法により推定されたフロンティア関数：シミュレーションデータに対して区分線形モデル（$m = 4$）を推定した。

誤差項 e_{ki} $(i = 1, \cdots, n_k;\ k = 1, \cdots, m)$ を推定された区分線形方程式の残差 $e_{ki} = \hat{a}(k)^{\mathrm{LS}} + \hat{b}(k)^{\mathrm{LS}} X_i - y_i$ $(X_i \in \mathrm{I}_k^{(n)})$ とし，この量は非負の値をとる。例えば損失関数を

$$L(a, b) = \sum_{k=1}^{m} \sum_{i=1}^{n_k} e_{ki} \tag{2.14}$$

として, 推定係数に制約を課したうえで，この損失を最小にする最適な m を求めることが考えられる。

なお他の損失関数を用いることもできる。残差が非負であるため，観測値と推定された効率的フロンティアとの隔たりを用いるのは自然である。しかし，観測値が明らかに非効率で境界曲線から大きく離れている場合には，これらの点を損失関数から無視してもよいだろう。この点は伝統的な DEA の方法に関連している可能性がある。また, 重複する区間をとることも考えられるが, この場合には異なる区間から求められる誤差は必ずしも独立とは限らないことに注意する必要があろう。

3　データ数が大きい場合のSDEA

3.1　極値DEA法

標本サイズが大きい場合には，回帰DEA法を改良することが可能であり, 統計的極値理論（SEVT）に基づく第2の推定法として極値DEA法を説明する。このSEVTは統計学の一分野として発展してきたもので，自然災害や金融危機のような極端で稀な事象の分析に威力を発揮するが, モーメントや正規分布に基づく標準的統計学では扱いにくい極端な現象を統計分析することから発展している[1]。SEVTでは極値分布として3類型が議論されているが，SDEA モデルでは観測値に上限が存在し，それが我々の統計解析の主対象であることから，第2（Weibull）型極値分布が自然に現れる。SEVT に極値DEA法では誤差項についてはモーメント条件は必ずしも必要としない。

区間 $\mathrm{I}_k^{(n)}$ （$k = 1, \cdots, m;\ m \geq 3$）をとり，$\mathrm{I}_k^{(n)} = (w_1^{(k)}, w_2^{(k)}]$ $(w_1^{(k)} < w_2^{(k)})$，$0 \leq w_1^{(1)} < \cdots < w_1^{(m)}$，$0 \leq w_2^{(1)} < \cdots < w_2^{(m)}$ とする。極値EDA法では，$\mathrm{I}_k^{(n)}$ $(k = 1, \cdots, m)$ から2つの区間 $\mathrm{I}_k^{(n)}(j)$ $(j = 1, 2)$ （簡単化して $\mathrm{I}(j)$ と記する）を選び，二つの区間のデータ数をそれぞれ $n(j)$ $(j = 1, 2)$, $n = n(1) + n(2)$ とする。（二つの区間は必ずしも隣接しなくともよい。）
$\bar{X}_L = (1/n(1)) \sum_{X_i \in \mathrm{I}(1)} X_i$，$\bar{X}_M = (1/n(2)) \sum_{X_i \in \mathrm{I}(2)} X_i$, $0 < \bar{X}_L < \bar{X}_M$ を仮定する。さらに二つの区間における Y の最大観測値をそれぞれ $Y_L(1) = \max_{X_i \in \mathrm{I}(1)} Y_i$，$Y_M(2) = \max_{X_i \in \mathrm{I}(2)} Y_i$ とする。そして極値DEA法による推定量を

$$\hat{b}(k) = \frac{Y_M(2) - Y_L(1)}{\bar{X}_M - \bar{X}_L} \tag{3.1}$$

および

$$\hat{a}(k) = \min_{X_i \in \mathrm{I}(1) \cup \mathrm{I}(2)} \{a \mid a + \hat{b}(k) X_i \geq Y_i\} \,. \tag{3.2}$$

により定める。なお, 係数の推定量に対し $\mathrm{I}_k^{(n)}$（$k = 1, \cdots, m$) で単調性制約

$$0 \leq \hat{a}(1) \leq \cdots \leq \hat{a}(m) \,, \quad \hat{b}(1) \geq \cdots \geq \hat{b}(m) \geq 0 \tag{3.3}$$

を課す。いずれかの区間で推定係数がこの必要条件を満たさない場合には，当該区間の推定値を利用せず、二つの区間を取り直せばよい。

次に推定量の漸近的性質を考察する為に次のように区間幅についての条件を考える。ここである正の数 c_j $(j = 1, 2)$ と $\delta (> 0)$ を選び

$$\text{(条件 } \mathbf{A}) \ \mathbf{I}_k^{(n)}(1) = \left\{ x \mid |x - \bar{X}_L| \leq \frac{c_1}{n(1)^\delta} \right\}, \ \mathbf{I}_k^{(n)}(2) = \left\{ x \mid |x - \bar{X}_M| \leq \frac{c_2}{n(2)^\delta} \right\}$$

[1]詳細は例えばEmbrechts, P., Klüppelberg, C. and Mikosch (1997), 高橋・志村 (2016) などがある。

と定める。
この条件について若干の補足しよう。X_i $(i=1,\cdots,n(j);\ j=1,2)$ について，2 つの区間で $n(j)$ に依存する $X_i^{(n)}$ をとり $n(j)$（$n=n(1)+n(2)$）に応じた $\mathrm{I}_k^{(n)}$ の列をとると，区間の長さ $(c_1+c_2)/n(j)^\delta$ $(\delta>0)$（$j=1,2$）は n の増加とともに短くなる。すなわち $n(j)\to\infty$ $(j=1,2)$ のもとで入力変数の二つの水準 $\bar{X}_L$, $\bar{X}_M$ の近傍に多くのデータが得られることを想定している。これは 3.2 節でより詳しく議論する統計的極値論と繰り返し観測が可能な場合を拡張したと見なすことができ，極値 EDA 条件と呼べるだろう。この条件は，区間内で $\bar{X}_L$ と $\bar{X}_M$ の近傍に十分なデータ点が存在することを意味する。
この極値 EDA 法による係数推定量 $\hat{a}(k)$ と $\hat{b}(k)$ の一致性について次の結果を得るが、証明は数学補論に与えておく。

定理 2： U_i (≤ 0) は i.i.d. な確率変数列であり，分布関数 F は密度 $f(u)$ をもち，$u=0$ で $f(u)$ は有界かつ滑らかとする。また X_i は有界で，$0<\bar{X}_L<\bar{X}_M$ を仮定する。(2.7), (3.1), (3.2）において $n\to\infty$（$n(1),n(2)\to\infty$）の場合を考え，$n/n(1)$ および $n/n(2)$ が正の定数に収束すると仮定する。このとき任意の $\delta(>0)$ について $n\to\infty$ のとき

$$\begin{bmatrix} \hat{a}(k)-a(k) \\ \hat{b}(k)-b(k) \end{bmatrix} \xrightarrow{p} \mathbf{0}\,. \tag{3.4}$$

となる。フロンティア関数を

$$\hat{h}_m(x)=\hat{a}(k)+\hat{b}(k)x \quad (\text{任意の } x\in \mathbf{I}_k^{(n)}) \tag{3.5}$$

により構成すれば，$\hat{h}_m(x)-h_m(x)=[\hat{a}(k)-a(k)]+[\hat{b}(k)-b(k)]x \xrightarrow{p} 0$ であるから, 区分線形関数 $h_m(x)$ の一致推定量が得られる。

次に推定量の漸近分布を考察する。極値 DEA 推定量の漸近的性質は δ の値に依存するが、導出は数理的補論に与える。

定理 3： U_i (≤ 0) は i.i.d. な確率変数列であり，分布関数 F は密度 $f(u)$ をもち，$u=0$ で $f(u)$ は（滑らかで）有界であるとする。また X_i は有界で，$0<\bar{X}_L<\bar{X}_M$ を仮定する。(2.7), (3.1), (3.2）において $n\to\infty$（$n(1),n(2)\to\infty$）の場合を考え，$n/n(1)$ および $n/n(2)$ が正の定数に収束すると仮定する。
(i) $0<\delta\le 1$ のとき任意の $0<\alpha<\delta$ に対し，

$$n^\alpha[\hat{b}(k)-b(k)] \xrightarrow{p} 0\,, \tag{3.6}$$

$$n^\alpha[\hat{a}(k)-a(k)] \xrightarrow{p} 0\,, \tag{3.7}$$

が $n\to\infty$ で成り立つ。
(ii) $\lambda_{1n}=[\frac{n}{n(1)}]\,[1/(\bar{X}_M-\bar{X}_L)]$，$\lambda_{2n}=[\frac{n}{n(2)}]\,[1/(\bar{X}_M-\bar{X}_L)]$ とおく。$\delta>1$ とする

と $n \to \infty$ のとき，

$$n[\hat{b}(k) - b(k)] \xrightarrow{w} Z_b = \lambda_2 Z_2 - \lambda_1 Z_1 \ , \tag{3.8}$$

ここで Z_i $(i = 1, 2)$ は互いに独立で，$G(\lambda) = e^{\lambda z_i}; (z_i \leq 0;\ i = 1, 2)$ に従い，$\lambda = f(0)$ である。Z_b の分布は $G_b(z) = [\lambda_2/(\lambda_1 + \lambda_2)] \exp[(\lambda/\lambda_2)z]$ $(z < 0)$，$G_b(z) = 1 - [\lambda_1/(\lambda_1 + \lambda_2)] \exp[-(\lambda/\lambda_1)z]$ $(z \geq 0)$ で与えられる。ただし $\lambda_1 = \lim_{n,n(2)\to\infty} \lambda_{1n}$，$\lambda_2 = \lim_{n,n(1)\to\infty} \lambda_{2n}$，$\lambda_1 > 0,\ \lambda_2 > 0$ の有限値に収束すると仮定する。
(iii) $\delta > 1$ のとき，$n \to \infty$ とすると，

$$n[\hat{a}(k) - a(k)] \xrightarrow{w} Z_a = c_{11} Z_1 - c_{12} Z_2 \ , \tag{3.9}$$

ここで $c_{11} = \lim_{n,n(1)\to\infty} [\, n/n(1) + \lambda_{1n} \bar{X}_L \,]$，$c_{12} = \lim_{n,n(2)\to\infty} [\, \lambda_{2n} \bar{X}_L \,]$ とし，$c_{11} > 0,\ c_{12} > 0$ の有限値に収束すると仮定する。
Z_a の分布は $G_a(z) = \frac{c_{11}}{c_{11}+c_{12}} \exp[(\lambda/c_{11})z]$ $(z < 0)$，$G_a(z) = 1 - \frac{c_{12}}{c_{11}+c_{12}} \exp[-(\lambda/c_{12})z]$ $(z \geq 0)$ で与えられる。

[注意 **1**] : データが与えられた時には δ を設定できる。δ を大きくとると選択した区間に存在するデータ数が少なくなるというトレード・オフが存在する。適切な選択を考察する必要があるが、$\delta \sim 1$ として漸近分布を利用することが考えられる。

[注意 **2**] : $G_b(z)$ の密度関数は

$$g_b(z) = \Big[\frac{\lambda}{\lambda_1 + \lambda_2}\Big] \exp\Big[\frac{\lambda}{\lambda_2} z\Big] \ (z < 0) \ , \quad g_b(z) = \Big[\frac{\lambda}{\lambda_1 + \lambda_2}\Big] \exp\Big[- \frac{\lambda}{\lambda_1} z\Big] \ (z \geq 0) \ , \tag{3.10}$$

$G_a(z)$ の密度関数（$c_{11} > 0,\ c_{12} > 0$）は

$$g_a(z) = \Big[\frac{\lambda}{c_{11} + c_{12}}\Big] \exp\Big[\frac{\lambda}{c_{11}} z\Big] \ (z < 0) \ , \quad g_a(z) = \Big[\frac{\lambda}{c_{11} + c_{12}}\Big] \exp\Big[-\frac{\lambda}{c_{12}} z\Big] \ (z \geq 0) \ . \tag{3.11}$$

$n(1) = n(2)$ のときには $\lambda_1 = \lambda_2$ となり，Z_b と Z_a の分布は両側指数分布（二重指数分布）になる。$\hat{b}(k)$ と $\hat{a}(k)$ の漸近分布における収束オーダーは $\sqrt{n}$ ではなく n である。これは，区間内の最大値の情報を効率的に用いようとする極値 DEA 法に基づいているためである。

区分線形関数 $h_m(x)$ に対して $X = x$ とおき，$n[\hat{a}(k) - a(k)]$ の極限を確率変数 Z_a により表現すれば, 漸近的表現が得られる。$x = \bar{X}$ が与えられるとき，$\hat{h}_m(x) - h_m(x) \xrightarrow{p} 0$ であり，極限の確率変数は

$$n[\hat{h}_m(x) - h_m(x)] \xrightarrow{w} Z_h = Z_a + Z_b x = (c_{11} - \lambda_1 x) Z_1 - (c_{12} - \lambda_2 x) Z_2$$

と表される。ここで $a_1 = c_{11} - \lambda_1 x = [n/n(1)][(\bar{X}_M - x)/(\bar{X}_M - \bar{X}_L) > 0$，$a_2 = c_{12} - \lambda_2 x = [n/n(2)][(\bar{X}_L - x)/(\bar{X}_M - \bar{X}_L) < 0$ であるから, Z_i $(i = 1, 2)$ が (負の)

指数分布にしたがうことを用いると，Z_h の分布関数は簡単な計算により,

$$\mathbf{P}(Z_h \le z) = \frac{a_1}{a_1+a_2}\exp\Big[\frac{\lambda}{a_1}z\Big] + \frac{a_2}{a_1+a_2}\exp\Big[-\frac{\lambda}{a_2}z\Big] \quad (z \le 0) \tag{3.12}$$

と与えられる。
この漸近分布は負値のみをとり, 未知パラメータ $\lambda(>0)$ に依存する。$\mathbf{I}(j)$ $(j=1,2)$ における残差 $\hat{U}_i = Y_i - \hat{a}(k) - \hat{b}(k)X_i$ を用いて λ を $(-1)\hat{\lambda}^{-1} = (1/n)\sum_{i=1}^{n}\hat{U}_i$ により推定するのが自然であり, λ の信頼区間を構成できる。

3.2 繰り返し観測の場合

本節では，繰り返し観測のケースを用いて，極値 DEA 法を（古典的）統計的極値理論（SEVT）と関連づける。ここでは簡単化ために $a(k)=0$ $(k=1,\cdots,m)$、非効率性項は未知の連続分布 F に従う i.i.d. 確率変数列であると仮定しておく。固定された X に対して繰り返し観測が得られる場合を考える。これは条件 A を満たす極限的なケースに対応する。X_k $(k=1,\cdots,m)$ を用いて

$$Y_{kj} = b(k)X_k + U_{kj} \quad (k=1,\cdots,m;\ j=1,\cdots,n_k) \tag{3.13}$$

と書く。ここで U_{kj} (≤ 0) は分布関数 F をもつ i.i.d. 確率変数列であり，切片は 0 とする。

説明変数の値 $X_k = x$ が与えられ，各区間で多数の観測値があり $n_k \to +\infty$ となる状況を考える。ここで F は 0 近傍で滑らかとする。このとき次の関係を利用する

$$\begin{aligned}
\mathrm{P}\Big(\max_{j=1,\cdots,n_k} Y_{kj} \le z_n\Big) &= \prod_{j=1}^{n_k}\mathrm{P}(U_{kj} \le z_n - b(k)X_k) \\
&= \exp\left\{\sum_{j=1}^{n_k}\log\Big[F[(z_n - b(k)X_k)\wedge 0]\Big]\right\} \\
&= \exp\left\{n_k \log\Big[1 - \frac{1}{n_k}\,n_k\,\bar{F}((z_n - b(k)X_k)\wedge 0)\Big]\right\},
\end{aligned}$$

ここで関数 $\bar{F}(x) = 1-F(x)$ とすると, $\bar{F}(x)$ は分布の右側テールであり，U_{kj} は非正の確率変数であることに注意しよう。$z_n = b(k)X_k + z/n_k$ $(z<0)$ とおく。すると，$\bar{F}(x)$ の $x=0$ まわりのテイラー展開（$\bar{F}(0)=0$，$F(0)-F(z/n_k) \sim f(0)\times[-z/n_k]$）を用いると，$k=1,\cdots,m$ について $n_k \to \infty$ のとき

$$\mathrm{P}\Big(n_k[\max_{j=1,\cdots,n_k} Y_{kj} - b(k)X_k] \le z\Big) \longrightarrow \exp[f(0)\,z] \quad (z<0) \tag{3.14}$$

となることが分かる。ただし $f(0)$ が有界であると仮定したが, 極限分布は負の指数分布 $F(u) = \exp[\lambda u]$ $(u \leq 0)$ であり，$\lambda(> 0)$ を用いて $f(0) = \lambda$ となる。

より一般には，(3.13) の非効率性項 U_{kj} の密度 $f(x)$ が $x = 0$ で発散する場合も考えられる。典型例は右端点が有限のパレート型分布で，0 近傍で $f(x) \sim C(-x)^{\alpha-1}$ $(x < 0,\ 1 > \alpha > 0)$ （ある定数 C に対して）となる場合である。一つの分布クラスとして，$y = -x$ (> 0) に対し

$$\bar{F}(-y^{-1}) = y^{-\alpha}L(y) \tag{3.15}$$

という形を仮定することができる。ここで $L(y)$ は緩慢変動関数 (slowley varying function) で $\alpha > 0$ である。（正の関数 L が $(0, \infty)$ 上で緩慢変動であるとは，任意の $t > 0$ に対し条件 $\lim_{y\to\infty}[L(ty)/L(y)] = 1$ を満たすことをいう。Embrechts et al. (1997) の p.564 を参照。）このとき Embrechts, P., Klüppelberg, C., Mikosch (1997) の定理 3.3.12 より，$c(n_k) = -F^{\leftarrow}(1 - n_k^{-1})$ を選ぶと，$n_k \to \infty$ のとき

$$\mathrm{P}\Big(c(n_k)^{-1}[\max_{j=1,\cdots,n_k} Y_{kj} - b(k)X_k] \leq z\Big) \longrightarrow \exp[-(-z)^{\alpha}] \ \ (z \leq 0)\,, \tag{3.16}$$

が成り立つ。ここで $\alpha > 0$，$F^{\leftarrow}(t) = \inf\{x \mid F(x) \geq t\}$（$0 < t < 1$）である。この定式化は統計的極値理論（SEVT）における標準的なもので，第 2（Weibull）型分布の 最大吸引域（maximum domain of attraction）として知られている。この漸近分布は第 2 型（Weibull）極値分布であるが[2]。この場合には一般にはスケール母数 α を推定する必要があり，これは必ずしも自明ではない。

SDEA の問題では，ノイズ（非効率性項）に対し，一般のケース (3.13) で $\alpha(> 0)$ を想定することも可能である。しかし本研究では，$z = 0$ における密度の有界性仮定が多くの応用で妥当と考えられることから，このより一般的な定式化は採用しなかった。企業の効率性の分析例では、通常は特定の産業で多くの企業が非効率であるが、他方、一部の企業がフロンティア関数の近傍に存在する状況で，DEA によりフロンティア関数を推定することに意味があるだろう

SDEA の設定では観測は多いことが多いが，本節で議論した繰り返し観測モデルと同一であるとは限らない。しかし繰り返し観測のケースに類似した状況は多く存在し，ある x に対して $[x - c_n,\, x + c_n]$ の領域に多数の観測が得られるような正数列 $c_n > 0$ をとることができる。3.1 節の 条件 A は，たとえば $c_n = c_j/n(j)^{\delta}$ $(j = 1, 2;\ \delta > 0)$ とおく場合に対応する。漸近理論を展開するためには，このほかにも様々なタイプの状況や条件を考えることができる。

[2]極値統計量が第 2 型極値分布に収束する一つの必要十分条件は高橋・志村 (2016),25 項が議論している。

4　推定法の一般化の考察

4.1　回帰 DEA 推定

実際の DEA の応用では説明変数が複数あることが一般的である。例えば 6 節の企業経営分析の入力は複数あるのが一般的である。そこで前節までの方法を一般化して説明変数の個数を p とする。ここでは説明のため $p=2$ の場合を用いるが, 領域の記法として $\mathbf{I}_k^{(n)}$（ただし $k=(k_1,k_2)$，$m=(m_1,m_2)$）を用い，$\mathbf{I}_k^{(n)}=\mathbf{I}_{1,k_1}^{(n)}\times\mathbf{I}_{2,k_2}^{(n)}=(w_1^{(k_1)},w_2^{(k_1)}]\times(w_1^{(k_2)},w_2^{(k_2)}]$ と定める（$k_1=1,\cdots,m_1$，$k_2=1,\cdots,m_2$）。このときフロンティア曲面を凹性制約の下で次の形の超平面

$$h_m(\mathbf{X})=a(k)+b_1(k)X_1+b_2(k)X_2 \tag{4.1}$$

を利用して推定する問題である。ただし $\mathbf{X}=(X_1,X_2)'\in\bigcup_k\mathbf{I}_k^{(n)}$ とする。
ベクトル $\mathbf{X}=(X_1,X_2)'$，$\mathbf{X}(i)=(X_1(i),X_2(i))'$，$\mathbf{X}(j)=(X_1(j),X_2(j))'$（$i\neq j$）が $\bigcup_k\mathbf{I}_k^{(m)}$ に属し，非負のスカラー λ_i,λ_j（$i\neq j$）をとるとする。このとき凹性制約は

$$h_m(\mathbf{x})\geq\lambda_i h_m(\mathbf{x}(i))+\lambda_j h_m(\mathbf{x}(j)) \tag{4.2}$$

を意味する。ただし $\mathbf{x}=\lambda_i\mathbf{x}(i)+\lambda_j\mathbf{x}(j)$，$\lambda_i+\lambda_j=1$ $(\lambda_i\geq 0,\ \lambda_j\geq 0)$ である。
これらの条件は各推定時に数値的に検証できるが，数値評価にはいくつかの注意が必要となる。例として $i=1,\ j=2$ の場合を考え，以下の手順を用いる。
(Step 1)：まず領域 $\mathbf{I}_k^{(n)}=\mathrm{I}_{1,k_1}^{(n)}\times\mathrm{I}_{2,k_2}^{(n)}$（$X_1\in\mathrm{I}_{1,k_1}^{(n)}$，$X_2\in\mathrm{I}_{2,k_2}^{(n)}$ において，全データを用い，係数に $a(k)\geq 0,\ b_1(k)\geq 0,\ b_2(k)\geq 0$ という制約を課して $h_1(X_1,X_2)=a(k)+b_1(k)X_1+b_2(k)X_2$ を推定する。
(Step 2)：次に $\mathbf{I}^{(n)}(1)$ とその近傍のいくつかの異なる区間 $\mathbf{I}_k^{(n)}$ $(k=2,\cdots,m)$ を取り，各領域で局所的に接平面 $h_1(X_1,X_2)=a(k)+b_1(k)X_1+b_2(k)X_2$ を推定し，凹性制約と係数の非負性を確認する。満たされない場合はその推定結果を棄却し，満たされる場合は推定結果と区分線形関数を採用する。
(Step 3)：同じ手順を繰り返す。データ点が有限であるため手続きは最終的に停止する。（数値実験では，標本サイズの 0.1 未満となるように m を取り，m_1 と m_2 を選べばよい。）

未知係数の推定を複数説明変数へと拡張した回帰 EDA 法をより具体的に説明しよう。係数ベクトルについて最小二乗推定を適用し，出力水準の調整によって切片係数を構成する。その後，単調性制約を満たすまで係数の構成を続ける。
第 2 節の回帰ベースの方法は，説明変数が複数の場合にも直ちに拡張できる。係数 $b_j(k)$ $(j=1,\cdots,p)$ は線形回帰式

$$\hat{\mathbf{b}}(k)^{LS}=\Big[\sum_{i=1}^{n}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})'\Big]^{-1}\Big[\sum_{i=1}^{n}(\mathbf{X}_i-\bar{\mathbf{X}})Y_i\Big] \tag{4.3}$$

により推定できる。ただし $\mathbf{X}_i = (X_{ji})$ は $p \times 1$ の入力ベクトル，$\bar{\mathbf{X}} = (1/n)\sum_{i=1}^{n} \mathbf{X}_i$（領域 $\mathbf{I}_k^{(m)}$ における平均点），Y_i は出力変数である。切片推定量は

$$\hat{a}(k)^{LS} = \min_{i=1,\cdots,n} \{a \mid a + \hat{\mathbf{b}}(k)^{LS\prime}\mathbf{X}_i \geq Y_i\} \tag{4.4}$$

で与えられる。$\mathbf{b}(k) = (b_j(k))$ と $a(k)$（$j = 1, \cdots, p;\ k = 1, \cdots, m$）の漸近分布の次数は $\sqrt{n}$ であり，定理 1 と系 1 を直接拡張できる。また $p \geq 1$ の一般の場合には $k = 1, \cdots, m$ について

$$Y_i = a(k) + \sum_{j=1}^{p} b_j(k) X_{ji} + U_i \ \ (i = 1, \cdots, n)\,, \tag{4.5}$$

（$U_i \leq 0$）と書ける。

このとき $\sqrt{n}\,[\hat{\mathbf{b}}(k) - \mathbf{b}(k)]$ の漸近分布は平均 0 の p 次元正規分布で，分散共分散行列は $\mathbf{AV} = \sigma_u^2 \mathbf{M}_x^{-1}$ である。ただし $\mathbf{M}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' \to \mathbf{M}_x$，かつ $\mathbf{M}_x$ は正則と仮定する必要がある。また定理 2 より $\sqrt{n}\,[\hat{a}(k) - a(k)]$ の極限分布はやや複雑でることがわかる。$\mathbf{b}^{**} = (b_j^{**})$ を $\frac{\mathbf{M}_n^{1/2}}{\sigma_u}\sqrt{n}\,[\hat{\mathbf{b}}(k)' - \mathbf{b}(k)']$ の極限の確率ベクトル, また $-\mathbf{X}_L^{*(n)} = (-X_{jL}^{*(n)}) = -\sigma_u \mathbf{M}_n^{-1/2}\mathbf{X}_L^{(n)}$，$-\mathbf{X}_M^{*(n)} = (-X_{jM}^{*(n)}) = -\sigma_u \mathbf{M}_n^{-1/2}\mathbf{X}_M^{(n)}$ を正規化した下限・上限の $p \times 1$ ベクトルとする（$\mathbf{I}_k^{(n)}$ における境界ベクトルを $\mathbf{X}_L^{(n)}, \mathbf{X}_M^{(n)}$ であるが, ここでは $\lim_{n\to\infty} \mathbf{X}_L^{*(n)} = \mathbf{X}_L^* = (X_{jL}^*)$，$\lim_{n\to\infty} \mathbf{X}_M^{*(n)} = \mathbf{X}_M^* = (X_{jM}^*)$ が存在すると仮定している)。$\sqrt{n}\,[\hat{a}(k) - a(k)]$ の極限の確率変数を a^* とすれば，$a_L^* \leq a^* \leq a_M^*$ であり，

$$a_M^* = \sum_{j=1}^{p} [\,-b_{j+}^* X_{jL}^* - b_{j-}^* X_{jM}^*], \qquad a_L^* = \sum_{j=1}^{p} [\,-b_{j+}^* X_{jM}^* - b_{j-}^* X_{jL}^*] \tag{4.6}$$

と表現できる。ただし $b_{j+}^{**} = b_j^{**}$ (if $b_j^{**} \geq 0$)，$b_{j-}^{**} = b_j^{**}$ (if $b_j^{**} < 0$) である。ここで確率変数 a_M^* と a_L^* は定理 1 で与えられるような区分多変量正規分布にしたがう確率変数である。したがって多くの場合には $p = 1$ の場合と同様に $X_{jc}^* = (X_{jL}^* + X_{jM}^*)/2 (j = 1, \cdots, p)$ により漸近分布を正規近似できる。

4.2 極値 DEA 推定

第 3 節で述べた SEVT に基づく凹性制約付きの第 2 の推定法を拡張するが, ここでは $p = 2$ の場合を説明する。ある k と m に対し領域 $\mathbf{I}_k^{(n)}$ を取り（n_{k_1,k_2} はデータ数；$k_1 = 1, \cdots, m_1$，$k_2 = 1, \cdots, m_2$），$\mathbf{I}(j, j') = \mathrm{I}_1(j) \times \mathrm{I}_2(j')$ $(j, j' = 1, 2)$ という 3 つの領域を用いる。各領域のデータ数を $n(j, j')$ とし，各領域の算術平均を $\bar{\mathbf{X}}(j, j') = (\bar{X}_1(j, j'), \bar{X}_2(j, j'))'$ $(j, j' = 1, 2)$ とする（$\bar{X}_1(j, j') = (1/n(j, j'))\sum_{X_i \in \mathbf{I}(j,j')} X_{1i}$，$\bar{X}_2(j, j') = (1/n(j, j'))\sum_{X_i \in \mathbf{I}(j,j')} X_{2i}$）。$p = 2$ のとき，$\bar{X}_1(1, 1) < \bar{X}_1(2, 1)$ か

つ $\bar{X}_2(1,1) < \bar{X}_2(2,1)$ を満たすように 3 領域がある。各領域における最大値は $Y_M(j,j') = \max_{\mathbf{X}_i \in \mathbf{I}(j,j')} Y_i \quad (j,j'=1,2)$ である。
このとき誤差項の 0 近傍での分布関数の滑らかさの仮定を用いると，まず

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

が成り立つ。
ここで正の数 $c > 0$ と $\delta > 0$ が存在して，$\mathbf{I}_k^{(n)}$ 内の列 $\mathbf{X}_i$ を $((j,k) = (2,1),(1,2),(1,1)$ に対し次の条件 A* を満たすように設定する：

$$\text{(条件 A}^*\text{)} \quad \mathbf{I}_k^{(n)}(j,k) = \{\, \mathbf{x} || x_{ji} - \bar{x}_j(j,k)| \le \frac{c}{n(j,k)^\delta}, |x_{ki} - \bar{x}_k(j,k)| \le \frac{c}{n(j,k)^\delta} \quad \}\,.$$

このように $p = 1$ における条件 A を拡張して漸近的正当化を考察する。3 節と同様の議論から $\max_{\mathbf{I}(2,1)} Y_i - \{a + b_1 \bar{X}_1(2,1) + b_2 \bar{X}_2(2,1)\} \xrightarrow{p} 0$, $\max_{\mathbf{I}(1,1)} Y_i - \{a + b_1 \bar{X}_1(1,1) + b_2 \bar{X}_2(1,1)\} \xrightarrow{p} 0$, $\max_{\mathbf{I}(1,2)} Y_i - \{a + b_1 \bar{X}_1(1,2) + b_2 \bar{X}_2(1,2)\} \xrightarrow{p} 0$ が成り立つ。したがって

$$[Y_M(2,1) - Y_M(1,1)] - b_1[\bar{X}_1(2,1) - \bar{X}_1(1,1)] - b_2[\bar{X}_2(2,1) - \bar{X}_2(1,1)] \xrightarrow{p} 0\,,$$

$$[Y_M(1,2) - Y_M(1,1)] - b_1[\bar{X}_1(1,2) - \bar{X}_1(1,1)] - b_2[\bar{X}_2(1,2) - \bar{X}_2(1,1)] \xrightarrow{p} 0$$

が得られる。そこで傾き係数 $(\hat{b}_1, \hat{b}_2)$ の推定量を

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

で定義する。切片推定量 $\hat{a}$ は

$$\hat{a} = \min_{\mathbf{X}_i \in \mathbf{I}(2,1) \cup \mathbf{I}(1,2) \cup \mathbf{I}(1,1)} \{a \mid a + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} \ge Y_i\} \tag{4.7}$$

により定義する。条件 A* の下では

$$\begin{bmatrix} \bar{X}_1(2,1) - \bar{X}_1(1,1) & \bar{X}_2(2,1) - \bar{X}_2(1,1) \\ \bar{X}_1(1,2) - \bar{X}_1(1,1) & \bar{X}_2(1,2) - \bar{X}_2(1,1) \end{bmatrix} \begin{bmatrix} \hat{b}_1 - b_1 \\ \hat{b}_2 - b_2 \end{bmatrix} \xrightarrow{p} 0$$

が成り立つ。ここで傾きの推定量が一致性を持つには非退化条件

$$\text{(条件 B)} \qquad \mathrm{rank} \begin{bmatrix} 1 & \bar{X}_1(2,1) & \bar{X}_2(2,1) \\ 1 & \bar{X}_1(1,2) & \bar{X}_2(1,2) \\ 1 & \bar{X}_1(1,1) & \bar{X}_2(1,1) \end{bmatrix} = 3$$

を仮定する。応用上では $p=2$ の領域を選択する際にこの条件を満たすことは可能と思われる。(もしデータ上で成り立たなければ領域の設定を変更する必要がある。) この非退化条件の下で，$\hat{b}_1 - b_1 \xrightarrow{p} 0$，$\hat{b}_2 - b_2 \xrightarrow{p} 0$ となる。さらに

$$Z_n(2,1) = n(2,1)[\max_{\mathbf{I}(2,1)} Y_i - \{a + b_1\bar{X}_1(2,1) + b_2\bar{X}_2(2,1)\}],$$

$$Z_n(1,2) = n(1,2)[\max_{\mathbf{I}(1,2)} Y_i - \{a + b_1\bar{X}_1(1,2) + b_2\bar{X}_2(1,2)\}],$$

$$Z_n(1,1) = n(1,1)[\max_{\mathbf{I}(1,1)} Y_i - \{a + b_1\bar{X}_1(1,1) + b_2\bar{X}_2(1,1)\}]$$

とおこう。条件 A* において $\delta > 1$ が選べるなら，極限として指数分布に従う確率変数 $Z(2,1), Z(1,2), Z(1,1)$ が得られ，その同時分布は

$$G(z_{21}, z_{12}, z_{11}) = \exp\left[\lambda(z_{21} + z_{12} + z_{11})\right] \quad (z_{21} \le 0,\ z_{12} \le 0,\ z_{11} \le 0)$$

である。したがって $n[\hat{b}_1 - b_1,\ \hat{b}_2 - b_2]'$ の漸近分布は指数分布の重み付き平均にしたがう確率変数ベクトル $\mathbf{b}^* = \mathbf{C}^*\mathbf{Z}_b$ の同時分布として与えられる。ただし

$$\mathbf{C}^* = \begin{bmatrix} \bar{X}_1(2,1) - \bar{X}_1(1,1) & \bar{X}_2(2,1) - \bar{X}_2(1,1) \\ \bar{X}_1(1,2) - \bar{X}_1(1,1) & \bar{X}_2(1,2) - \bar{X}_2(1,1) \end{bmatrix}^{-1} \begin{bmatrix} \lambda(2,1) & 0 & -\lambda(1,1) \\ 0 & \lambda(1,2) & -\lambda(1,1) \end{bmatrix},$$

$$\mathbf{Z}_b = \begin{bmatrix} Z(2,1) \\ Z(1,2) \\ Z(1,1) \end{bmatrix},\ \bar{\mathbf{X}}(j,k) = \begin{bmatrix} \bar{X}_1(j,k) \\ \bar{X}_2(j,k) \end{bmatrix}\ ((j,k) = (2,1),(1,2),(1,1))\ ,$$

および $\lambda(2,1) = \lim_{n\to\infty} n/n(2,1) > 0$，$\lambda(1,2) = \lim_{n\to\infty} n/n(1,2) > 0$，$\lambda(1,1) = \lim_{n\to\infty} n/n(1,1) > 0$ とした。すなわち $n(\hat{b}_j - b_j)$ $(j=1,2)$ の漸近分布は正規分布とはならず, 定理 3 を一般化した加重指数分布にしたがうことが分かる。このことから $n(\hat{a}-a)$ の漸近分布は指数分布にしたがう確率変数ベクトル $\mathbf{Z}_b = (Z(2,1), Z(1,2), Z(1,1))'$ がとる領域により表現できる。
またフロンティア関数については $n[\hat{h}_m(\mathbf{x}) - h_m(x)] \xrightarrow{w} Z_h$（ただし $\mathbf{z}_b = (z_{21}, z_{12}, z_{11})'$, $\mathbf{x} = (x_1, x_2)'$) とすると

$$P(Z_h \le z) = \int_{\mathbf{A}_{3n}} \lambda^3 \exp[\lambda(z_{21} + z_{12} + z_{11})] dz_{21} dz_{12} dz_{11}\ . \tag{4.8}$$

ただし, 積分領域は

$$\mathbf{A}_{3n} = \bigcap_{j,k} \{(\frac{n}{n(j,k)}) z_{jk} + (\mathbf{x} - \bar{X}(j,k)')\mathbf{C}^*\mathbf{z}_b \le z\}\ ,$$

$((j,k) = (2,1),(1,2),(1,1))$ で与えられる。
なおここで導出された極限分布の表現 (4.8) は $p=1$ の場合に得られた (3.12) の拡張になっている。

5　シミュレーション実験

SDEA 解析についていくつかのシミュレーションを行った。計算プログラムを開発、多くのシミュレーションを行ったが、ここでは (i) $p = 1$（説明変数が 1 つ）と (ii) $p = 2$（説明変数が 2 つ）の 2 つの場合の回帰 DEA による分析結果、(iii) データ数が多い場合の極値 DEA による分析結果を報告する。

回帰 EDA を使った実験では「再帰的な 2 分割を行う方法」と「区間の数 K を指定する方法」の両方を実装した。前者の方法では (2.14) に基づく計算結果の分布データサイズ N を 300、500、2000、5000 と指定し、それぞれについて 1000 回のリサンプリングを行い、サンプルから式 (2.14) に基づく計算結果を検討した。さらに各データセットに対して後者の方法を適用、区間の数（K）はアルゴリズムがさらに 2 分割できなくなるまで自動的に決定した。このとき、同じデータ生成過程で作られた同じデータサイズのサンプルであっても、リサンプリングごとに K の値は異なる場合がある。すべての K に対して式 (2.14) を計算し、それらをヒストグラムで検討、各 K に対する (2.14) 式の結果は全体として正規分布にしたがつていると見なせる。

シミュレーション実験では区間の取り方により、例えば区間にデータが存在しない、あるいはデータ数が少なく推定値が不安定になることや、制約条件を満たさないことが生じた。そこで推定に際しては理論と整合的になるように幾つかの工夫を行った。一次元の場合、区間境界を t_k $(k = 1, \cdots, m)$ とすると推定する区分線形関数が境界値で連続となるように $a(k)+b(k)t_k = a(k+1)+b(k+1)t_k$ $(k = 1, \cdots, m-1)$ という制約をかけた推定法が有効であることが分かった。

第 1 例では $h(X) = X^d$，第 2 例では $h(X_1, X_2) = X_1^d X_2^{1-d}$，$d = 0.3$ とする。両場合で説明変数 X は $N(0, 1)$ から，$-U$ は $\mathrm{Exp}(\lambda)$（$\lambda = 1/10$）から生成した。第 1 例では 200 個のデータを生成し，回帰ベースの方法で係数を推定した。その後，(2.14) を用いて最適な区間数を探索した。初めは再帰的手順を用い，まず 2 区間で線形回帰を推定し，各区間をさらに 2 つに分割して 2 本の回帰直線を当てはめることを繰り返し，基準関数が増加するまで続けた。しかし場合によっては，さらに探索を続ければ改善の余地があるにもかかわらず手続きが停止することが分かった。そこで素朴な方法に切り替え，まず m を固定して (2.14) を計算し，全体で評価することにした。

第 1 例では，(2.14) を再評価した結果，$m(= m^*) = 39$ で最小となることが分かった（図 3 参照)。m が大きい場合，(2.14) の値に不安定さが見られることがある。図 4 は当てはめた回帰直線と，推定したフロンティア関数は赤で示した。第 2 例では 400 個のデータを生成し，$e_{k,i}$ $(i = 1, \cdots, n_{k_1,\cdots,k_p};\ k_j = 1, \cdots, m_j;\ j = 1, \cdots, p)$ を推定された区分線形方程式の残差

$$e_{ki} = \hat{a}(k) + \hat{b}_1(k)X_{1i} + \hat{b}_2(k)X_{2i} - y_i \quad (\mathbf{X}_i = (X_{1i}, X_{2i})' \in \mathbf{I}_k^{(n)})$$

とし（非負値），$\mathbf{b}(k) = (b_1(k), b_2(k))'$ に対する損失関数を

$$L(a, \mathbf{b}, \mathbf{m}, p) = \sum_k \sum_{\mathbf{x}_i \in \mathbf{I}_k^{(m)}} e_{ki} \tag{5.1}$$

と定めた。そのうえで，推定係数に（本文の表記に従い）凸性制約を課しつつ損失を最小にする m を選んだ。最小値は $(m_1^*, m_2^*) = (4, 6)$ で達成された（図 5 参照）。$p = 2$ の場合は推定すべきパラメータが増えるため，主としてその理由で損失関数の変動が大きくなる。$\mathbf{m} = (m_1, m_2)'$ の選択は，単一説明変数の場合に比べて難しい。図 6 に推定したフロンティア関数を示しておく。

次にデータ数がかなり多い場合のシミュレーションの一例として n=6,000, m=35 に設定して極値 DEA 法を利用して分析した結果を報告する。シミュレーション・モデルは (i) と同様であるが、図 7 により例示しているが極値 DEA 法により真のフロンティア関数をかなりうまく推定していることを確認した。この場合には回帰 EDA 法よりも極値 DEA 法の方が推定効率が良いことも確認される。さらに図 7 におけるノイズは負の指数分布から発生させたが、図 8 ではデータのより多くがフロンティア関数に近く、ノイズ項は混合指数分布にしたがうシミュレーションの結果を示しておく。この場合、極値 EDA 法の推定精度が非常に良いという興味深い実験結果が得られた。

なお、今回のシミュレーション例はかなり極端な場合を想定した例であり、二つの推定法（回帰 EDA と極値 EDA）のメリット・デメリットについてはなお検討の余地があろう。

6　分析例：日本の生命保険業

本節では実証例として，前節までに説明した SDEA の方法日本の生命保険業の会計データに適用する。ＯＲ分野における EDA の適用対象は広範囲であるが, 基本的には入力に対して効率的な経営状態の評価という経済・経営における重要な意味を持っている。ここで紹介する分析は企業マクロ・データに基づく例示である。

ここで用いるデータは『生命保険事業概況』（生命保険協会（2021））に掲載された 2017〜2021 会計年度の公開データである。分析で用いた変数は，(1) works : 従業員数（事務職員），(2) capital : 株主資本合計，(3) expense : 経常事業費，(4) insurance : 保険金等支払金合計，(5) income : 経常利益であり，出力変数は経常利益（ordinary income）とした。

当該産業の企業数は 41 社と少ないため，フロンティア関数の推定には回帰 DEA を用いた。図 9 に示すとおり，見かけ上，フロンティア関数に関する単調性・凹性の仮定が満たされていない。したがって DEA 分析からは「かんぽ生命」を外れ値とし

て扱い，除去するのが適切と考えられる。41 社のうち，かんぽ生命は歴史的背景や制度変更のため他社と大きく異なり、郵便事業の一部分として運営してきた事情がある。なお日本の生命保険業は大きな相互会社が存在し続けている, という意味で民間企業としては典型的とは言えない。また日本における生命保険業の歴史的発展は，英国や米国などの生命保険業と比較すると, 他の先進諸国とは後発であり相当に異なっている。特に日本で主要な生命保険会社の数が比較的少ないことには歴史的・制度的理由がある。かんぽ生命はもともと日本の郵政事業の一部であり，2006 年に民営化された。生命保険・損害保険産業の歴史的展開の詳細については 久保（2011）を参照されたい。

データ分析では最終的に 40 社のデータに焦点を当てた。入力として従業員数，出力として経常利益を用い，2021 年度のフロンティア関数を図 10 に示しておくが、入力として資本，出力として経常利益を用いた場合、企業規模の大小に関わらず妥当なフロンティア関数が推定されている。また入力として資本，出力として経常利益を用いた場合の 2021 年度のフロンティア関数を図 11 に示しておく。これら 2 つの図から，フロンティア関数は概ね妥当に推定されていることが分かる。すなわち，推定されたフロンティア関数に近い企業もあれば，非効率な企業も存在することが確認できる。なお生命保険業には大規模企業がごく少数しかないため，右側（大規模側）領域でのフロンティア関数の推定は統計的に難しい問題であることが分かった。

さらに，説明変数として従業員数と資本の 2 変数（$p = 2$）を用いた場合のフロンティア関数を推定した。しかしデータ数が少ないため，局所的に選んだ領域によっては最小二乗推定が計算できないケースが生じる。領域にデータが存在しないような極端な場合には最小二乗推定は可能ではない。今回のデータ分析では 18 ケースについては計算可能であり，その結果として得られた損失関数の値を図 12 に示しておく。残差の分析により、最適値として $(m_1^*, m_2^*) = (2, 2)$ を選択したので、図 13 に日本の生命保険業における推定したフロンティア関数を示しておく。

7 結語

本研究では DEA における統計的課題を取り上げ，線形回帰と極値分布に基づく SDEA における 2 つの推定法を提案した。これらの統計的推定法はオペレーションズ・リサーチおよび統計学の双方にとって新規性があると考えられる。特に本稿で提案している方法ではフロンティア関数については効率性を表現する凹関数は仮定するが、微分可能性などその他の数理的条件は仮定していないことを強調しておく。ここで提案する回帰 DEA 法は，撹乱に対する 2 次モーメント条件を仮定する回帰分析の直接的応用であるのに対し，極値 DEA 法は誤差項のモーメント条件を必要ないが, 誤差項の端点付近の挙動が重要となることが分かった。本稿では原点付近での密度関数の有界性を利用して分析を行ったが、より一般化することは可能だろう。

こうした非効率性を示す分布の裾に関する情報を有効に活用することはORや計量経済学の問題ではこれまでそれほど注目されていないようであるが、新たなデータ分析の視角を提供する可能性があると思われる。

また応用例として，日本の生命保険業のデータ分析例を報告した。本例ではデータ量がかなり小さいため，係数推定には回帰ベースの手法を用いたが, データが豊富な場合には，追加条件の下で SEVT を用いる極値 DEA 法により効率性を向上できる可能性があるが、実用的な意味があるか否かは今後の課題とする。

ここで提案した統計的 DEA に関連して今後検討すべき課題が少なくない。第一に，本研究で与えた極限分布の精度の問題である。例えば極値 EDA 法の漸近理論の適用には条件 A を用いたが，有限標本での妥当性は重要な研究テーマである。第二に，本稿で用いた統計モデルは 4 節で説明したように多変量の入力・出力などいくつかの方向に拡張可能である。例えば SEV 理論では，最大値だけでなく上位順序統計量を併用する推定法があり，フロンティア関数 (包絡線) に関する情報を含みうる。ただし有限のデータからフロンティア関数を推定する際に単調性・凹性制約を課すときにどの様な情報が利用可能になるかは自明でない。データが十分に大きければ，説明変数をより効率的に利用できる可能性がある。有限データにおける入力変数数や区間（または領域）数の選択も重要であり，ランダムに領域を選ぶ場合も含め検討課題である。

最後に，本稿で議論した 統計的 DEA の問題設定は，凸集合内の有限データから接超平面を利用してフロンティア関数 (包絡曲線）を推定するという幾何学的にはより一般的な問題の特殊例と解釈できることを指摘しておく。本稿ではこの問題に対する二つの統計的推定法を提案したが, ここで開発した方法は，さらに他の状況へ一般化する可能性がある。

参考文献

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

[5] Kunitomo, N. and Zhao, Y. (2023),“A Statistical Data Envelopment Analysis”, 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.

[6] 久保英也 8(2011) 「日本の保険会社における経営統合効果の計測」保険学雑誌, 平成 23 年 3 月,179-198, 保険学会.

[7] Jirak, M., Meister, A. and Reiss, M. (2014), “Adaptive Function Estimation in Nonparametric Regression With One-sided Errors,,” *The Annals of Statistics*, 42-5, 1970-2002.

[8] Mas-Collel, A., M. Whinston and J. Green (1995), *Microeconomic Theory*, Oxford University Press.

[9] Millimet, D., and Parmeter, C. (2021), “Accounting for Skewed or One-Sided Measurement Error in the Dependent Variable,,” *Political Analysis*, 1-23.

[10] 高橋倫也・志村隆彰 (2016), 「極値統計学」, 近代科学社.

[11] 生命保険事業概況（2021）（日本語），生命保険協会（Life Insurance Association of Japan）.

数学的補論：定理の証明

この数理的補論では，前節までに示した定理の導出を与える。

定理 1 の証明 : 前半は線形回帰の標準的な議論を用いる。(2.7) と (2.8) を用いると，

$$\sqrt{n}[\hat{b}(k)^{LS} - b(k)]) = \frac{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(X_j - \bar{X})(U_j - \bar{U})}{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2} \ . \tag{A.1}$$

$\mathbf{E}[U_j] \leq 0$ ではあるが，$U_j - \bar{U} = (U_j - \mathbf{E}[U_j]) - (\bar{U} - \mathbf{E}[U_j])$ （$j = 1, \cdots, n$）と分解できるので，回帰分析では U_j の代わりに $U_j - \mathbf{E}[U_j]$ に標準的議論を適用できる。(A.1) の分母は M_x に収束し，分子は中心極限定理（CLT）により分布収束して $N(0, \sigma_u^2 M_x)$ に収束する。
次に，関係式

$$\hat{a}(k)^{LS} - a(k) \ = \ \min_{i=1,\cdots,n}\{a \mid Y_i \leq a + \hat{b}(k)^{LS} X_i\} - a(k)$$

$$= \max_{i=1,\cdots,n}\{Y_i - \hat{b}(k)^{LS}X_i\} - a(k)$$
$$= \max_{i=1,\cdots,n}\{U_i + [b(k) - \hat{b}(k)^{LS}]X_i\}$$

を用いる。$U_i \leq 0$ $(i = 1, \cdots, n)$ であるから，任意の $i = 1, \cdots, n$ について

$$U_i + [b(k) - \hat{b}(k)^{LS}]X_i \leq \max_{i=1,\cdots,n}\{U_i + [b(k) - \hat{b}(k)^{LS}]X_i\} \leq \max_{i=1,\cdots,n}\{[b(k) - \hat{b}(k)^{LS}]X_i\}$$

が成り立つ。したがって

$$\begin{aligned}\max_{i=1,\cdots,n}\{\sqrt{n}U_i\} + \sqrt{n}[b(k) - \hat{b}(k)^{LS}]X_{i^{*(n)}} &\leq \sqrt{n}[\hat{a}(k)^{LS} - a(k)] \\ &\leq \max_{i=1,\cdots,n}\{\sqrt{n}[b(k) - \hat{b}(k)^{LS}]X_i\}\ ,\end{aligned}$$

ここで $i^{*(n)}$ は $i = 1, \cdots, n$ に依存する確率変数である。
さらに $\max_{i=1,\cdots,n}\{U_i\} = O_p(1/n)$ なので，$\sqrt{n}\max_{i=1,\cdots,n}\{U_i\} \xrightarrow{p} 0$。したがって X_i $(i = 1, \cdots, n)$ が有界であるとき，$\max_{i=1,\cdots,n}\{\sqrt{n}[b(k) - \hat{b}(k)^{LS}]X_i\}$ の漸近分布を考える。$b^* = b^*_+ + b^*_-$ であり，$0 < X_L \leq X_i \leq X_M$ $(i = 1, \cdots, n)$ から $\max\{-b^*X_i\} \leq -b^*_+X_L - b^*_-X_M = a^*_M$ を得る。
$z \geq 0$ のとき，

$$\begin{aligned}\mathrm{P}(-b^*_+X_L - b^*_-X_M \leq z) &= \mathrm{P}(b^* \geq 0, b^* \geq -\frac{z}{X_L}) + \mathrm{P}(b^* < 0, b^* \geq -\frac{z}{X_M}) \\ &= 1 - \mathrm{P}(b^* \leq -\frac{z}{X_M})\ .\end{aligned}$$

同様に $z < 0$ では

$$\mathrm{P}(-b^*_+X_L - b^*_-X_M \leq z) = 1 - \mathrm{P}(b^* \leq -\frac{z}{X_L})$$

を得る。これにより a^*_M の密度関数が得られる。同様に，

$$\sqrt{n}[\hat{b}(k)^{LS} - b(k)]_+[-X^{(n)}_L] + \sqrt{n}[\hat{b}(k)^{LS} - b(k)]_-[-X^{(n)}_M] \leq \sqrt{n}[b(k) - \hat{b}(k)^{LS}]X_{i^{*(n)}} \quad \text{(A.2)}$$

を任意の $i^{*(n)}$ について用いると，左辺の極限分布は $b^*_+[-X_M] + b^*_-[-X_L]$ により抑えられることが分かる。よって a^*_L の分布が得られる。
(**Q**.**E**.**D**.)

定理 2 の証明 : z_n と $X_i \in \mathrm{I}(1)$ に対して，

$$\begin{aligned}\text{(A.3)}\mathrm{P}(\max_{X_i\in\mathrm{I}(1)} Y_i \leq z_n) &= \prod_{i=1}^{n(1)} \mathrm{P}\left(Y_i \leq z_n\right) \\ &= \prod_{i=1}^{n(1)} \mathrm{P}\left(Y_i - (a(k) + b(k)X_i) \leq z_n - (a(k) + b(k)X_i)\right) \\ &= \prod_{i=1}^{n(1)} \mathrm{P}\left(U_i \leq z_n - (a(k) + b(k)X_i)\right)\end{aligned}$$

が成り立つ。$\delta > 0$ なので $0 < \alpha < \delta$ をとる。$z_n = a(k) + b(k)\bar{X}_L + z/n(1)^\alpha$ $(z < 0)$ とおけば，この確率は

$$\prod_{i=1}^{n(1)} F\left(\left[\frac{z}{n(1)^\alpha} + b(k)(\bar{X}_L - X_i)\right] \wedge 0\right) \tag{A.4}$$

と書ける。条件 A の下では $\delta > \alpha > 0$，$1 > \alpha > 0$ をとると, 任意の $z < 0$ に対し，$z_n = [a(k) + b(k)\bar{X}_L] + \frac{z}{n(1)^\alpha}$，かつ $\alpha > \alpha' > 0$ とできる。すると (A.3) 右辺の支配項は

$$\mathrm{P}\left(\max_{X_i \in \mathrm{I}(1)} Y_i - (a(k) + b(k)\bar{X}_L) \leq \frac{z}{n(1)^\alpha}\right) \leq \exp\left[n(1) \log F\left(\frac{z}{n(1)^{\alpha'}}\right)\right] \to 0$$

$(n(1) \to \infty)$ となる。これは $\log F(z/n(1)^{\alpha'}) \sim \log[1 + f(0)\, z/n(1)^{\alpha'}] \sim f(0)\, z/n(1)^{\alpha'}$ $(0 < \alpha' < 1)$ による。
また任意の $\epsilon > 0$, $z < -\epsilon < 0$ に対して $\mathrm{P}\Big(\max_{X_i \in \mathrm{I}(1)} Y_i - (a(k) + b(k)\bar{X}_L) > \epsilon\Big) \to 0$ となる。
したがって，密度 f が 0 近傍で有界かつ滑らかであれば

$$\max_{X_i \in \mathrm{I}(1)} Y_i - (a(k) + b(k)\bar{X}_L) \xrightarrow{p} 0 \tag{A.5}$$

が成り立つ。
同様の議論を区間 $\mathrm{I}(2)$ と $\bar{X}_M$ $(> \bar{X}_L)$ に適用すると，$\max_{X_i \in I(2)} Y_i - (a(k) + b(k)\bar{X}_M) \xrightarrow{p} 0$ および

$$[\max_{X_i \in \mathrm{I}(2)} Y_i - \max_{i \in \mathrm{I}(1)} Y_i] - b_k[\bar{X}_M - \bar{X}_L] \xrightarrow{p} 0 \tag{A.6}$$

を得る。$0 < \bar{X}_L < \bar{X}_M$ から

$$\hat{b}(k) - b(k) \xrightarrow{p} 0 \tag{A.7}$$

が従う。パラメータ a については

$$\begin{aligned} \max_{X_i \in \mathrm{I}(1) \cup \mathrm{I}(2)} [Y_i - \hat{b}(k) X_i] &= \max_{X_i \in \mathrm{I}(1) \cup \mathrm{I}(2)} [a(k) + b(k) X_i + U_i - \hat{b}(k) X_i] \\ &= a_k + \max_{X_i \in \mathrm{I}_1 \cup \mathrm{I}(2)} [U_i + (b(k) - \hat{b}(k)) X_i] \end{aligned} \tag{A.8}$$

および

$$\mathrm{P}\left(\max_{X_i \in \mathrm{I}(1) \cup \mathrm{I}(2)} [Y_i - \hat{b}(k) X_i] - a(k) \leq z_n\right) = \mathrm{P}\left(\max_{X_i \in \mathrm{I}(1) \cup \mathrm{I}(2)} [U_i + (b(k) - \hat{b}(k)) X_i] \leq z_n\right)$$

を用いる。$b(k) - \hat{b}(k) \xrightarrow{p} 0$ かつ X_i 有界なので，$\epsilon_n = K/n^{1-\alpha}$ $(1 > \alpha > 0)$ を適当な $K > 0$ で取り，任意の K に対し $\mathrm{P}(|(b(k) - \hat{b}(k)) X_i| \leq \epsilon_n) \to 1$ とできる。

$z_n^* = z_n + \epsilon_n$（すなわち $z_n = z_n^* - \epsilon_n$）とおいて，定理 1 の証明の最後の議論と同様すると

$$\hat{a}(k) - a(k) \xrightarrow{p} 0 \tag{A.9}$$

を得る。（**Q.E.D.**）

定理 3 の証明 ： (i) $0 < \delta \leq$ の場合を考察する。まず $n^\alpha[\hat{b}(k) - b(k)]$ について定理 2 の証明を利用する。$0 < \alpha < \delta \leq 1$ に対し，$z_n = a(k) + b(k) + z/n^\delta$ とおくと $n^\alpha[\max_{\mathrm{I}(1)} Y_i - (a(k) + b(k)\bar{X}_L)] \xrightarrow{p} 0$ より $n^\alpha(\hat{b}(k) - b(k)) \xrightarrow{p} 0$ となる。
次に $\hat{a}(k) - a(k)$ の漸近的性質は，定理 1 の証明の最後と同様に議論により導ける。基準化した $(\hat{b}(k) - b(k))$ に極限については $0 < \alpha < 1$ かつ任意の負の z に対して $\mathrm{P}(n^\alpha(\hat{a}(k) - a(k)) \leq z) \to 0$ （$n \to \infty$）が成り立つ。(ただし $n(1)/n$ と $n(2)/n$ が正の定数に収束することを仮定した。)
(ii) 条件 A の下で $\delta > 1$ の場合を考える。$\hat{b}_k$ の漸近分布のため，$Z_{1n} = n(1)[\max_{\mathrm{I}_1} Y_i - (a(k) + b(k)\bar{X}_L)]$，$Z_{2n} = n(2)[\max_{\mathrm{I}_2} Y_i - (a(k) + b(k)\bar{X}_M)]$ とおく。すると

$$n[\hat{b}(k) - b(k)] = \frac{n}{\bar{X}_M - \bar{X}_L}\left[\frac{Z_{2n}}{n(2)} - \frac{Z_{1n}}{n(1)}\right] = \lambda_{2n} Z_{2n} - \lambda_{1n} Z_{1n} \tag{A.10}$$

となる。ここで $\lambda_{1n} = \frac{n}{n(1)(\bar{X}_M - \bar{X}_L)}$，$\lambda_{2n} = \frac{n}{n(2)(\bar{X}_M - \bar{X}_L)}$。
定理 2 と同様にして $\alpha = 1$ のときには $\exp[n(1)\log F(z/n(1))] \sim \exp[n(1)\log(1 + f(0)\,(z/n(1)))] \to \exp[\lambda z]$ $(z < 0)$ であるが，これは $F(0) = 1$，F は 0 近傍で密度 $f(z)$ をもつ滑らかな分布，$\lambda = f(0)$ としたことによる。Z_{1n} と Z_{3n} は独立なので，(Z_{1n}, Z_{2n}) の同時極限分布は $G(z_1, z_2) = \exp[\lambda(z_1 + z_2)]$ $(z_1 \leq 0, z_2 \leq 0)$ となる。
$\lambda_1 = \lim_{n\to\infty} \lambda_{1n}$，$\lambda_2 = \lim_{n\to\infty} \lambda_{2n}$ とおけば，極限の確率変数 $Z_b = \lambda_2 Z_{2n} - \lambda_1 Z_{1n}$ の分布が導ける。$Z_1 \leq 0$，$Z_2 \leq 0$ で $Z_b = \lambda_2 Z_2 - \lambda_1 Z_1$ は正負いずれの値も取り得るので注意が必要である。$Z_b \geq 0$ の場合，$\{Z_b \leq z\}$ かつ $Z_2 \leq 0$ から $(\lambda_2 Z_2 - z)/\lambda_1 \leq Z_1 \leq (\lambda_2/\lambda_1)Z_2$ を得る。$Z_b \leq 0$ の場合，$\{Z_b \leq z\}$ かつ $Z_1 \leq 0$ から $Z_2 \leq (\lambda_1 Z_1 + z)/\lambda_3$ となる。したがって場合分けが必要である。
$z < 0$ のとき

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

$z \geq 0$ のとき

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

$$
\begin{aligned}
&= \int_{-\infty}^{0} \lambda \exp(\lambda z_2)\Big\{ \exp(\lambda(\lambda_2/\lambda_1)z_2) - \exp(\lambda((\lambda_2 z_2 - z)/\lambda_1))\Big\} dz_2 \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2}\Big[1 - \exp\Big(-\frac{\lambda}{\lambda_1}z\Big)\Big].
\end{aligned}
$$

(iii) $G_a(z)$ の導出は $n[\hat{b}(k) - b(k)]$ の場合と同様である。次の関係を用いる:

$$
\begin{aligned}
&\text{(A.11)} \mathrm{P}\left(n[\max_{X_i \in \mathrm{I}(1)\cup\mathrm{I}(2)} (Y_i - \hat{b}(k)X_i) - a(k)] \le z \right) \\
&= \mathrm{P}\left(\max\Big\{ n[\max_{X_i\in\mathrm{I}(1)} (U_i + (b(k)-\hat{b}(k))X_i)],\, n[\max_{X_i\in\mathrm{I}(2)} (U_i + (b(k)-\hat{b}(k))X_i)] \Big\} \le z \right).
\end{aligned}
$$

条件 A かつ δ (> 1) の下で，$n\max_{X_i\in \mathrm{I}_1}[U_i + (b(k) - \hat{b}(k))X_i] = n\max_{X_i\in\mathrm{I}_1}[U_i + (b(k)-\hat{b}(k))\bar{X}_L] + o_p(1)$，同様に $n\max_{X_i\in\mathrm{I}_2}[U_i + (b(k)-\hat{b}(k))X_i] = n\max_{X_i\in\mathrm{I}_2}[U_i + (b(k)-\hat{b}(k))\bar{X}_M] + o_p(1)$ が成り立つ。
$c_{11,n} = [\frac{n}{n(1)} + \lambda_{1n}\bar{X}_L]$，$c_{22,n} = [\frac{n}{n(2)} - \lambda_{2n}\bar{X}_M]$，$c_{12,n} = \lambda_{2n}\bar{X}_L$，$c_{21,n} = \lambda_{1n}\bar{X}_M$ とおく。$\lambda_{1n} = [n/n(1)]/[\bar{X}_M - \bar{X}_L]$ であるから $c_{11,n} = c_{21,n}$，$c_{12,n} = c_{22,n}$ を得る。このとき (A.10) の極限で現れる確率変数を用いて，(A.11) は漸近的に

$$
\mathrm{P}(\max\{c_{11}Z_1 - c_{12}Z_2,\, c_{21}Z_1 - c_{22}Z_2\} \le z) = \mathrm{P}(c_{11}Z_1 - c_{12}Z_2 \le z)
$$

と等価になる。ここで $c_{ij} = \lim_{n\to\infty} c_{ij,n}$ $(i,j = 1,2)$。
最後の等式は $c_{11}Z_1 - c_{12}Z_3 = c_{21}Z_1 - c_{22}Z_2$ による。$Z_a = c_{11}Z_1 - c_{12}Z_2$ を $n(\hat{a}_k - a_k)$ の極限の確率変数とすると，Z_a は正負のいずれの値も取り得る。その分布関数は次のように導ける。
$z \le 0$,$Z_1 \le (z + c_{12}z_2)/c_{11} \le 0$ （$c_{11} > 0 > c_{12}$）のとき，

$$
\begin{aligned}
G_a(z) &= \int_{-\infty}^{0}\int_{-\infty}^{(z+c_{12}z_2)/c_{11}} \lambda^2 \exp[\lambda(z_1+z_2)]\, dz_1\, dz_2 \\
&= \int_{-\infty}^{0} \lambda \exp\Big(\lambda\frac{z + c_{12}z_2}{c_{11}}\Big)\, dz_2 \\
&= \frac{c_{11}}{c_{11}+c_{12}} \exp\left[\frac{\lambda}{c_{11}}z\right].
\end{aligned}
$$

$z > 0$, $z_1 \le 0$ かつ $(c_{11}z_1 - z)/c_{12} \le Z_2 \le 0$ のとき，

$$
\begin{aligned}
G_a(z) &= \int_{-\infty}^{0}\int_{(c_{11}z_1 - z)/c_{12}}^{0} \lambda^2 \exp[\lambda(z_1+z_2)]\, dz_2\, dz_1 \\
&= \int_{-\infty}^{0} \lambda \exp[\lambda z_1]\Big\{1 - \exp\left(\lambda(c_{11}z_1 - z)/c_{12}\right)\Big\} dz_1
\end{aligned}
$$

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

以上により定理 3(i)〜(iii) が得られる。
(**Q**.**E**.**D**.)

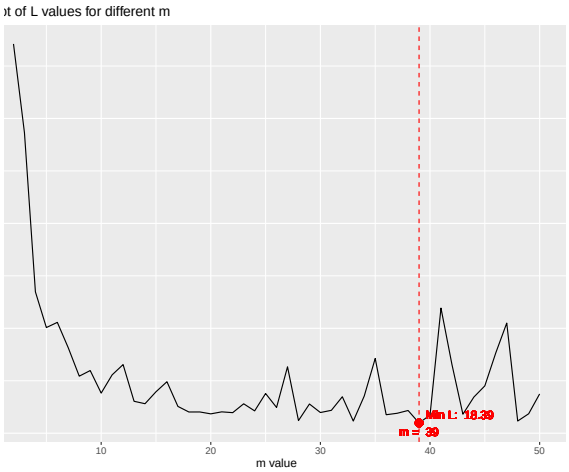

図 3: m の選択：区間数と (2.14) の関係：「再帰的な 2 分割を行う方法」により N=200 のデータに対して、誤差を評価した結果。区間を決定するときに重なりを含めるように設定、K を 2 から５０まで変化させたが、K=39 で最小の誤差 18.39 となった。

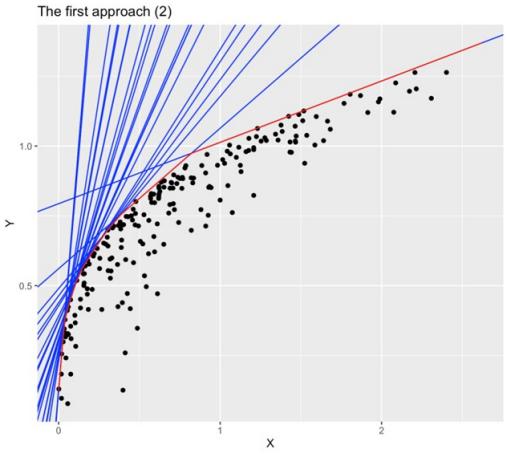

図 4: $m = 39$ のときの推定直線群のプロット

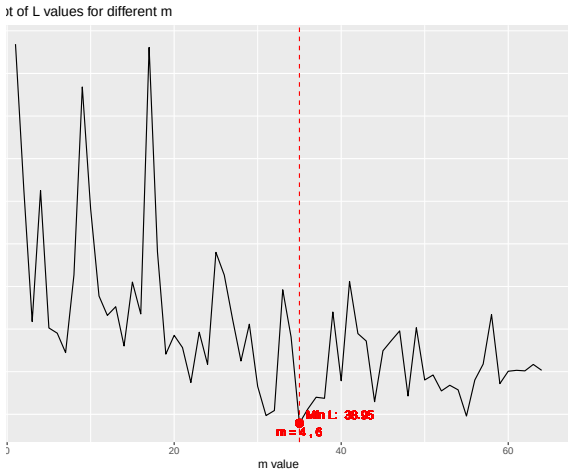

図 5: $\mathbf{m} = (m_1, m_2)$ の選択：領域数と (5.1) の関係：2 変数をもつ 400 個のデータを生成、(5.1) の誤差評価の式を用いて最適な区間を決定した。シミュレーションでは、分割 $a = 2, 3, 4, 5, 6, 7, 8, 9$、$b = 2, 3, 4, 5, 6, 7, 8, 9$ より合計 64 個の組み合わせを検討 $a = 4$、$b = 6$ のときに、誤差が最小 38.95 となった。

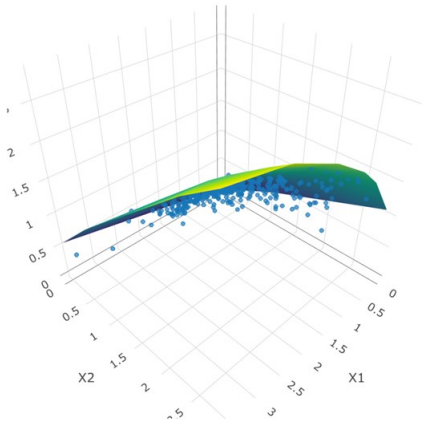

図 6: $(m_1, m_2) = (4, 6)$ のときの推定曲面：$a = 4$、$b = 6$ をのときのフロンティア曲面のプロット。3D プロットのため、一つの角度のスクリーンショット。

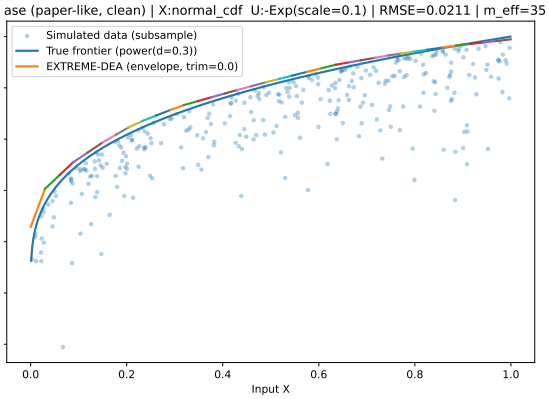

図 7: 極値 DEA によるフロンティア推定の基本設定として, 真のフロンティアはべき乗関数 $h(x) = x^{0.3}$, 入力 X は標準正規分布の累積分布関数により (0,1) 区間に変換して生成した。非効率項 U はスケールパラメータ 0.1 の指数分布の符号を反転させたもので、すべての観測値がフロンティア以下に位置する。標本サイズ N=6000、区間分割数 m=35 のもとで、推定フロンティア（カラーの区分線形線）が真のフロンティア (黒の曲線) に良く近似しており、RMSE=0.0211 と高い精度を示している。

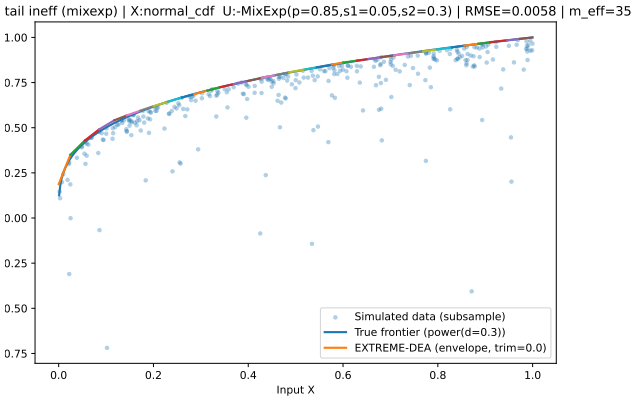

図 8: 非効率項の分布を 2 つの指数分布の混合とした場合の極値 DEA 推定である $(N = 6000, m = 35)$。具体的には、確率 0.85 でスケール 0.05 の指数分布（フロンティア近傍に集中する成分）、確率 0.15 でスケール 0.30 の指数分布（フロンティアから遠方に離れる成分）の混合分布を用いた。全体の 85 の観測値がフロンティアのごく近くに集中するため、上側包絡面の推定に有利な条件となり、RMSE=0.0058 と基本設定の場合を上回る高い精度が得られている。

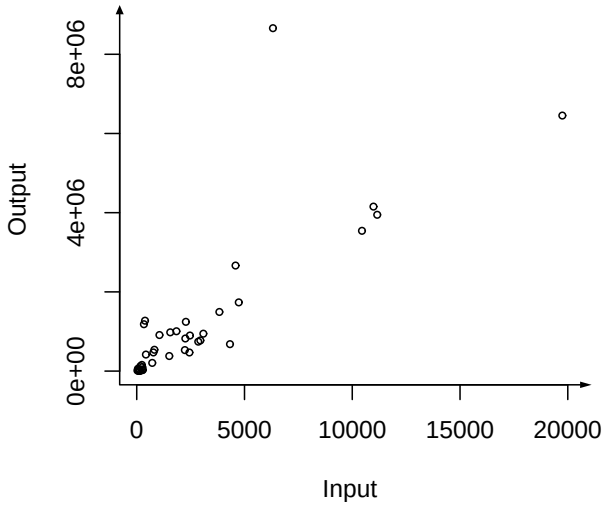

図 9: 外れ値の状況：日本の生命保険業では，外れ値と考えられる企業（かんぽ生命保険）がある。

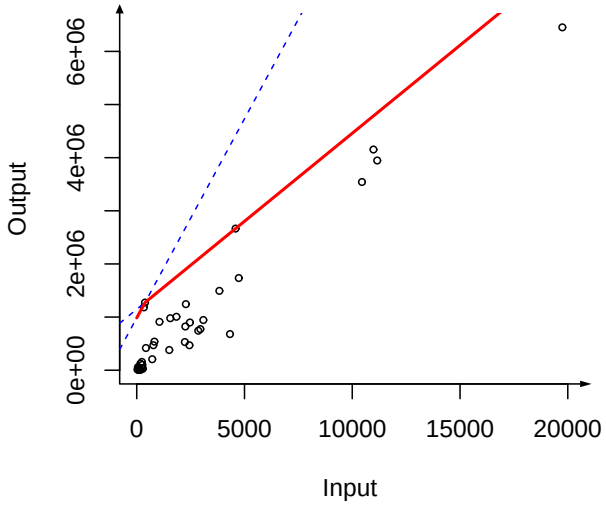

図 10: 推定されたフロンティア日本の生命保険業）：入力＝従業員数，出力＝経常利益, として回帰 DEA で推定した結果。

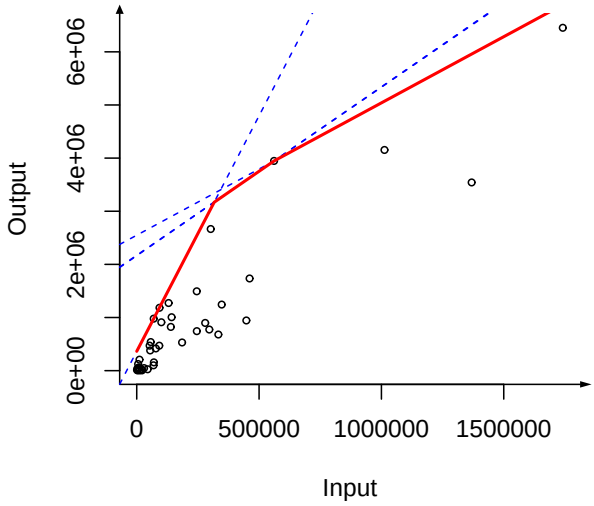

図 11: 推定されたフロンティア（日本の生命保険業）：入力＝資本，出力＝経常利益, として回帰 DEA で推定した結果。

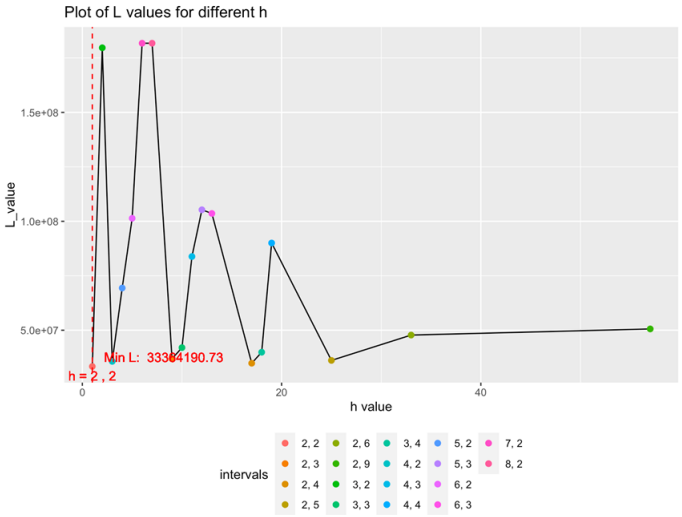

図 12: 2 つの入力変数 (従業員数と資本, 日本の生命保険業) を用いた場合の m の選択結果: 資本 (capital) と労働 (Labor) に対し、それぞれ K=2: 9 を指定し、区間の分割を行った。最適な分割は (2,2), 誤差値 33364190.73 となった。全部で $8 \times\ 8 = 64$ パターンがありうるが、データサイズが小さいため、分割により最小二乗法の推定に十分なデータが得られない区間にはその場合を除外した。

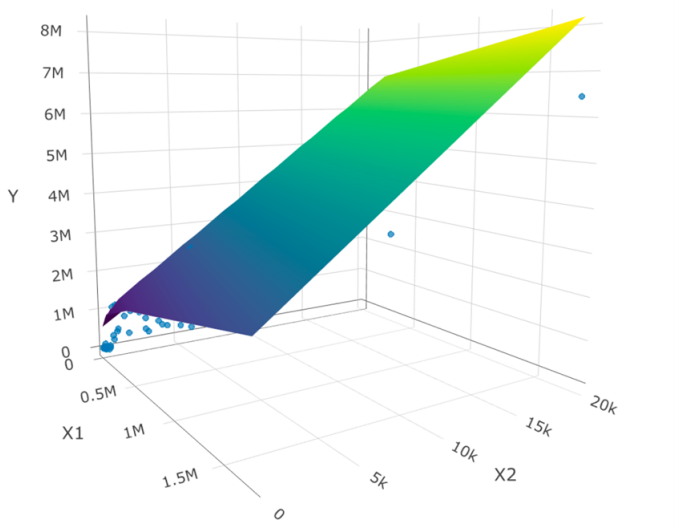

図 13: 2 つの入力変数 (従業員数と資本) を用いた場合の日本の生命保険業の推定されたフロンティア曲面：3D プロットのため、一つの角度のスクリーンショット。