統計的DEA :理論と応用

国友直人　大学統計教員養成センター 特任教授

1　DEAとは？

OR(operations research) ではData Envelopment Analysis （DEA, データ包絡分析）は企業の効率性評価の一般的な方法として知られている。伝統的には数理計画問題として定式化され様々な方法が開発されている (Cooper, Seiford, and Tone (2007))。DEA(Data Envelopment Analysis) は興味深いが, 確率的メカニズムを全く考えていない。現実の生産インプット・アウトプットには様々な原因から偶然性が含まれていると考えられる。そこで出力情報に対して入力情報を利用する統計モデルを考えるが、要点は観察される出力データは効率的フロンティア関数を超えない点にある。このDEAとは全く独立に計量経済学 (econometrics) における生産関数の計測問題ではパラメトリックな切断分布を用いたAigner, Lovell, and Schmidt (1977, Journal of Econometrics) などの方法も知られている。

これら既存の方法と異なる統計的DEA(SDEA) を開発, 応用中であるので報告する。この研究はZhao Yu氏(趙宇氏, 東京理科大学経営学部助教) との(統計エキスパート人材育成プロジェクトでの)共同研究である。

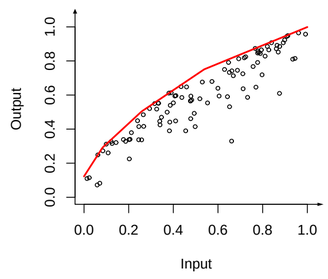

図 1: 推定された区分線形フロンティア関数

2　統計的区分線形DEA

入力(説明)変数Xの情報が利用可能なとき, まずXを固定して係数a, bを用いた線形関数を$\mu = a + bX$ を上界, 出力の最大値が利用できる状況を考える。局所的にXが観察できるとき, 出力データはYには非正値の誤差があり, n個$(n \geq 2)$の一次元観測データ$Y_1, Y_2, \cdots, Y_n$が$Y_i = a + bX_i + U_i \ \ (i = 1, \cdots, n)$ に従うとしよう。a, bは未知母数、X_iは説明変数, $U_i \leq 0$はi.i.d. 確率変数列, 分布関数をG(密度g)とする。ここで説明変数Xが存在すると, 観測値Yの最大値と想定される誤差項Uの最小値は必ずしも一致しない。非正値をとる誤差項とともに、説明変数Xがあるので, 既存の統計的極値論（例えばEmbrechts et al. (1997)）をそのまま適用できない。母数の推定には色々な方法が考えられるが, ここでは二つの方法を考察する。

[方法**1**] 効率フロンティアをまずデータから傾き母数bを最小二乗法を利用$\hat{b}_{LS}$として, aはデータを含む最大切片$\hat{a}_{LS} = \min_{i=1,\cdots,2n}\{a | a + \hat{b}_{LS}X_i \geq Y_i\}$ により推定する。

[方法**2**] インプットXの順序により2つの区間の各々の最大値を利用する。例えばデータ数を$2n$としてXの値域により2分割, 小さい方から$I_{(L)}, I_{(M)}$, 各区間でのXの平均を$X_{(L)}, X_{(M)}$ とする。各区間での最大値を$Y_{(L)}, Y_{(M)}$ とする。(i)まずb を$\hat{b}_{EV} = \frac{Y_{(M)} - Y_{(L)}}{X_{(M)} - X_{(L)}}$ により推定, (ii)次にaを$\hat{a}_{EV} = \min_{i=1,\cdots,2n}\{a | a + \hat{b}X_i \geq Y_i\}$ により推定する。

すべてのデータが推定された直線の下に来る条件下で最適化するのはより一般には「凸集合内の2n個の観測データから超平面(hyper-plane)を推定する」という統計的問題と解釈できる。データの上位値を利用するなど極値を利用するというアイデアは自然であるが統計的性質は自明ではなさそうである。いずれにしても母数aの推定には$\hat{a} = \max_{i=1,\cdots,2n}\{Y_i - \hat{b}X_i\}$ が良いと思われる。データ数$n \longrightarrow \infty$のとき一定の条件下で推定量が一致性を持つことが望ましいが, 漸近分布の形は知られていないと思われる。また上の方法でない他の良い推定方法もありそうである。

実証分析を考慮すると, 説明変数Xによりデータを分割して推定する方法が実際的である。いま説明変数を順序づけてm個の点を選び$w_{(1)} < w_{(2)} < \cdots < w_{(m)}$ とする。変数Xの領域を有界$(0 < X < M)$としてm個の区間$(m = m_n)$を$I_j \ (j = 1, \cdots, m)$ とする(I_jを重なっても良いが、各中心は単調増加とする)。各I_jにおいて$Y_i = \alpha_j + b_j X_i + U_i \ (X_i \in I_j, j = 1, \cdots, m)$ 未知母数$(a_j, b_j) \ j = 1, \cdots, m$ の推定を行うと, 「各区間内のデータ数$n_j \ (j = 1, \cdots, m)$が大きく、かつ区間数m を大きくとる」場合には正当化できよう。各セルでのデータ数$n_j \longrightarrow \infty$のとき$\begin{bmatrix} \hat{a}_j \\ \hat{b}_j \end{bmatrix} \xrightarrow{p} \begin{bmatrix} a_j \\ b_j \end{bmatrix} \ (j = 1, \cdots, m)$ が予想される。この定式化は統計的に自然であるが、この種の議論は既存の文献では見かけなさそうである。漸近分布が求まれば未知係数の不確実性を評価できる。ただし分布や説明変数の挙動について幾つかの条件が必要だろう。説明変数によりデータをセルに分割, 最大値を利用する場合, たとえ説明変数が有界としてもセル数が多くなるので, 各セル当たりのデータ数が小さくならないように領域をとる必要がある。各セルにはかなりのデータ数が必要となるのでm, n_jに関する正当化の理論が必要となる。データがそれほど多くない場合にはとりあえず方法1の適用が考えられる。

3　一般の統計的DEA問題

問題は一般化が可能である。(i) 説明変数Xが複数ある場合には$\mathbf{X}_i = (X_{1i}, \cdots, X_{Ki})$ に対して各変数の順序付けにより領域を構成することが考えられる。例えば$K = 2$の場合には二つの変数の順序付けを$w_{(11)} < w_{(12)} < \cdots < w_{(1m)}$, $w_{(21)} < w_{(22)} < \cdots < w_{(2m)}$ 行い, m^2個の領域$I_{ij} \ (i, j = 1, \cdots, m)$ に分割、例えば4領域における極値を利用することが考えられる。(ii) n個$(n \geq 2)$の一次元観測データ$Y_1, Y_2, \cdots, Y_n$を1変数Xで順序づけた$w_{(1)} < w_{(2)} < \cdots < w_{(m)}$ から校正した区間$I_j \ (j = 1, \cdots, m)$に対して$Y_i = f(X_i, \theta) + U_i \ \ (i \in I_j \ ; j = 1, \cdots, m)$ という非線形モデルが考えられる。ここでθは未知母数, X_iは説明変数, $U_i \leq 0$はi.i.d. 確率変数列, 分布関数をGとする。この時にはミクロ経済学の考察などからfは凹関数(concave)関数$f(\cdot, \cdot)$の制約があるので, 1次元なら条件$b_1 > \cdots > b_m \geq 0$が必要となる。効率的関数の推定は区分線形(piece-wise linear)関数$f(X, \theta) = a_j + b_j X \ (X \in I_j; \ j = 1, \cdots, m)$ の推定に還元することで簡単化される。各区間I_jにおけるデータ数n_jが多く, 区間数$m = m_n \to \infty$のときには任意の凹関数の効率曲線 f の推定が可能となるので, 統計的問題として興味深い。

我々の研究ではDEAに関する新しい統計的アプローチを検討している。SDEA(統計的DEA)法はフロンティア推定を効率的に近いと思われる観測データを利用する方法であり, 線形回帰分析や統計的極値論を応用して数理的正当化が可能と思われる。特定のパラメトリック・モデルを仮定しないので企業の効率性分析や生産関数のデータ分析に有用であろう。シミュレーションやデータ上での検証や簡単な実例として日本の生命保険業の分析を行ったが、なお理論と応用を含め課題が少なくない(なおKunitomo-Zhao(未発表DP,2022)があるが、含まれている誤りの修正を含め改訂中であることに注意されたい。SDEAについてのコメントは募集中である)。

参考文献

[1] 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.

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

[3] Kunitomo, N. and Zhao, Y. (2022), " A Statistical Data Envelopment Analysis," https://stat-expert.ism.ac.jp/training/discussionpaper/