区間関数による線形回帰

研究ノート

乾　口　雅　弘*

Linear Regression by Interval Function

Key Words：Interval function, interval data, possibility, necessity, Minkowski difference

1．はじめに

不確実性といえば，確率を連想される人が多い．しかし，非確率的な不確実性あるいは，確率の規則に従わない不確実性モデルも存在する．1990年代に家電製品などでブームとなったファジィ理論がそのようなモデルの一つである．著者は非確率的な不確実性の下での意思決定支援に役立つ手法について研究している．特にファジィ情報の下での最適化や識別不能性に基づくデータ解析法を与えるラフ集合などの研究が主である．本稿では，回帰問題を取り上げ，従来法とは異なった可能性概念に基づく区間回帰分析 [1, 2] を紹介する．

2．バラツキを可能性でとらえる

非確率的な不確実性を扱うモデルの大半は，与えられた起こりうる範囲から導かれる事象の可能性と必然性という互いに双対な二つの様相概念で取り扱われる．ここで取り上げる区間回帰分析では，説明変数に応じて定められる目的変数の取りうる範囲を定めようとするものである．

通常の回帰分析では，目的変数 y と説明変数ベクトル $\boldsymbol{x}$ との間に $y=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}+a_0$ が成立すると仮定し，与えられたデータ $\{(y_i, \boldsymbol{x}_i), i=1, 2, \ldots, n\}$ に見合ったパラメータ $\boldsymbol{a}$，a_0 が求められる．すなわち，y_i と $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i+a_0$ とのズレ $\varepsilon_i=|y_i-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i-a_0|$ は誤差と考えられ，この総和あるいは二乗総和が最小になるように $\boldsymbol{a}, a_0$ が推定される．

一方，区間回帰分析 [1] では，説明変数ベクトル $\boldsymbol{x}$ から目的変数 y が厳密に一つの値に定められるほど関係が精密でなく，$\boldsymbol{x}$ から y の取りうる範囲 $Y(\boldsymbol{x})$ が推定できるに過ぎないと考える．すなわち，$\boldsymbol{x}$ と y の関係自体が不明確さを伴っていると考え，区間関数 $Y(\boldsymbol{x})=\sum_{j=1}^{m} A_j x_j + A_0$ により $Y(\boldsymbol{x})$ を推定する．ただし，$\boldsymbol{x}=(x_1, x_2, \ldots, x_m)^{\mathrm{T}}$ であり，$A_i=[a_i^{\mathrm{L}}, a_i^{\mathrm{R}}]$, $i=1, 2, \ldots, m$ である．区間 A_j を中心 $a_j^{\mathrm{C}}=(a_j^{\mathrm{L}}+a_j^{\mathrm{R}})/2$ と幅 $a_j^{\mathrm{W}}=a_j^{\mathrm{R}}-a_j^{\mathrm{L}}$ を用いて，$A_j=\langle a_j^{\mathrm{C}}, a_j^{\mathrm{W}}\rangle$ と表記すれば，次式が得られる．

$$Y(\boldsymbol{x})=\left\langle \sum_{j=1}^{m} a_j^{\mathrm{C}} x_j + a_0^{\mathrm{C}}, \sum_{j=1}^{m} a_j^{\mathrm{W}} |x_j| + a_0^{\mathrm{W}} \right\rangle \quad (1)$$

データ $\{(y_i, \boldsymbol{x}_i), i=1, 2, \ldots, n\}$ が与えられたとき，$Y(\boldsymbol{x})$ が目的変数変数 y の取りうる範囲を示すためには，$y_i \in Y(\boldsymbol{x}_i)$, $i=1, 2, \ldots, n$ を満たす必要がある．この条件を満たす関数は無限に存在するので，$Y(\boldsymbol{x}_i)$ の幅の総和 $\sum_{i=1}^{n}\sum_{j=1}^{m} a_j^{\mathrm{W}}|x_j| + a_0^{\mathrm{W}}$ を最小にするように，$a_j^{\mathrm{C}}, a_j^{\mathrm{W}}$, $j=0, 1, \ldots, m$ が定められる（図1参照）．ただし，$\boldsymbol{x}_i=(x_{i1}, x_{i2}, \ldots, x_{im})^{\mathrm{T}}$,

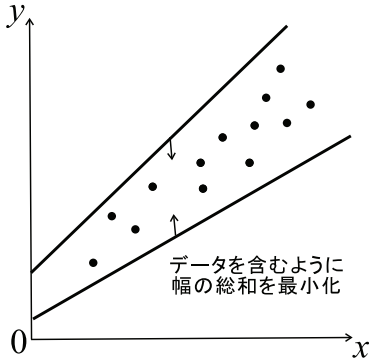


図1: 通常データの区間回帰

* Masahiro INUIGUCHI
1962年10月生
大阪府立大学　大学院工学研究科　博士前期課程　経営工学専攻修了（1987年）
現在、大阪大学　大学院　基礎工学研究科　システム創成専攻　教授　博士（工学）　システム計画数理
TEL：06-6850-6350
FAX：06-6850-6350
E-mail：inuiguti@sys.es.osaka-u.ac.jp

$i = 1, 2, \ldots, n$ である．この問題は線形計画問題となり，比較的容易に解を求めることができる．

3．区間データに対しては二つのモデル

官能検査，感性工学，意思決定などにおいて嗜好度や選好度などの評価モデルを作成する際には，説明変数ベクトル $\boldsymbol{x}_i$ に対する目的変数の値を区間値 $Y_i = [y_i^{\mathrm{L}}, y_i^{\mathrm{R}}]$ で測定した方が望ましい場合がある．また，考慮していない要因により説明変数ベクトル $\boldsymbol{x}_i$ に対する目的変数の値がある範囲 Y_i で変動するような場合も考えられる．これらの場合，区間データ $\{(Y_i, \boldsymbol{x}_i),\ i = 1, 2, \ldots, n\}$ を扱うことになる．

区間データに対しても，区間関数による回帰分析がいくつか考えられる．まず，与えられた Y_i，$i = 1, 2, \ldots, n$ と理論式 $Y(\boldsymbol{x})$ とから起こりうる最小の範囲を推定することが考えられる．この場合，$Y_i \subseteq Y(\boldsymbol{x}_i)$，$i = 1, 2, \ldots, n$ を満たす範囲で，幅の総和を最小にする $a_j^{\mathrm{C}}, a_j^{\mathrm{W}}$，$j = 0, 1, \ldots, m$ を求めればよいことになる（図2参照）．この問題は次の線形計画問題に帰着され，得られた $Y(\boldsymbol{x})$ は可能性モデル [1] と呼ばれる．

$$
\begin{aligned}
&\text{minimize} \quad \sum_{i=1}^{n}\sum_{j=0}^{m} a_j^{\mathrm{W}} |x_{ij}| \\
&\text{subject to} \\
&\sum_{j=0}^{m} a_j^{\mathrm{C}} x_{ij} - \frac{1}{2}\left(\sum_{j=0}^{m} a_j^{\mathrm{W}} |x_{ij}|\right) \le y_i^{\mathrm{L}}, \\
&\sum_{j=0}^{m} a_j^{\mathrm{C}} x_{ij} + \frac{1}{2}\left(\sum_{j=0}^{m} a_j^{\mathrm{W}} |x_{ij}|\right) \ge y_i^{\mathrm{R}}, \\
&\qquad\qquad i = 1, 2, \ldots, n \\
&a_j^{\mathrm{W}} \ge 0,\ j = 0, 1, \ldots, m
\end{aligned}
\tag{2}
$$

ここで，便宜上，$x_{i0} = 1$, $i = 1, 2, \ldots, n$ と定めている．以降も同様である．

一方，与えられた Y_i，$i = 1, 2, \ldots, n$ と理論式 $Y(\boldsymbol{x})$ とから間違いなく起こると考えられる最大の範囲を推定することも考えられる．この場合，$Y_i \supseteq Y(\boldsymbol{x}_i)$，$i = 1, 2, \ldots, n$ を満たす範囲で，幅の総和を最大にする a_j^{C}，a_j^{W}，$j = 0, 1, \ldots, m$ を求めればよいことになる（図2参照）．この問題は次の線形計画問題に帰着され，得られた $Y(\boldsymbol{x})$ は必然性モデル [1] と呼ばれる．

$$
\begin{aligned}
&\text{maximize} \quad \sum_{i=1}^{n}\sum_{j=0}^{m} a_j^{\mathrm{W}} |x_{ij}| \\
&\text{subject to} \\
&\sum_{j=0}^{m} a_j^{\mathrm{C}} x_{ij} - \frac{1}{2}\left(\sum_{j=0}^{m} a_j^{\mathrm{W}} |x_{ij}|\right) \ge y_i^{\mathrm{L}}, \\
&\sum_{j=0}^{m} a_j^{\mathrm{C}} x_{ij} + \frac{1}{2}\left(\sum_{j=1}^{m} a_j^{\mathrm{W}} |x_{ij}|\right) \le y_i^{\mathrm{R}}, \\
&\qquad\qquad i = 1, 2, \ldots, n \\
&a_j^{\mathrm{W}} \ge 0,\ j = 0, 1, \ldots, m
\end{aligned}
\tag{3}
$$

可能性モデルは常に存在するが，必然性モデルは存在するとは限らないことに注意しよう．

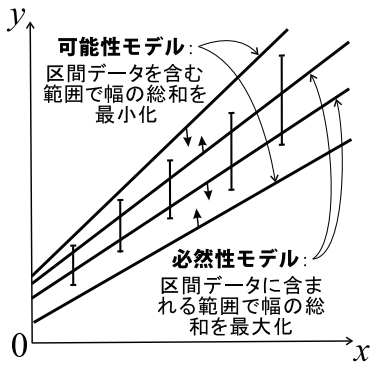


図2: 可能性モデルと必然性モデル

4．誤差最小化で扱えないか？

区間回帰分析では，通常の回帰分析と異なり，評価関数に誤差最小化という概念を用いていないが，誤差最小化という観点では同様な手法が得られないのであろうか？　そこで，観測された Y_i は理論区間 $Y(\boldsymbol{x}_i)$ に区間誤差 E_i^{L} が加わって得られていると考える．すなわち，

$$
Y(\boldsymbol{x}_i) + E_i^{\mathrm{L}} = Y_i, i = 1, 2, \ldots, n \tag{4}
$$

とし，次の区間絶対誤差和最小化問題を考える [2]．

$$
\text{minimize} \sum_{i=1}^{n} |E_i^{\mathrm{L}}| \tag{5}
$$

ここで，$E_i^{\mathrm{L}} = [e_i^{\mathrm{L}}, e_i^{\mathrm{R}}]$ とすると，その絶対値 $|E_i^{\mathrm{L}}|$ は次式で与えられる．

$$
\begin{aligned}
|E_i^{\mathrm{L}}| &= [\varepsilon_i^{\mathrm{L}}, \varepsilon_i^{\mathrm{R}}] \\
&= \begin{cases}
[\max(e_i^{\mathrm{L}}, -e_i^{\mathrm{R}}), \max(-e_i^{\mathrm{L}}, e_i^{\mathrm{R}})], \\
\qquad\qquad e_i^{\mathrm{L}} \cdot e_i^{\mathrm{R}} \ge 0 \\
[0, \max(-e_i^{\mathrm{L}}, e_i^{\mathrm{R}})], \quad e_i^{\mathrm{L}} \cdot e_i^{\mathrm{R}} < 0
\end{cases}
\end{aligned}
\tag{6}
$$

問題 (5) は区間値の最小化問題で，その意味が明確ではない．このような場合，区間値の最小化に対してある解釈を導入して問題を扱うことになる．ここでは，問題 (5) を次の問題として解釈する．

$$\text{lex-min}\left(\sum_{i=1}^{n}\varepsilon_i^{\mathrm{L}},\ \sum_{i=1}^{n}\varepsilon_i^{\mathrm{R}}\right) \tag{7}$$

lex-min は辞書式最小化を意味し，まず第1成分の評価関数（第1評価関数）を最小化し，それが最適値を取るという条件の下で第2成分の評価関数（第2評価関数）を最小化していくことを表す．この問題は2段階線形計画問題に帰着でき [2]，たとえば，シンプレックス法により解くことができる．これにより得られた $Y(\boldsymbol{x})$ はミンコフスキー差モデルと呼ばれている．

この問題に対して，次式が成立する．

$$\varepsilon_i^{\mathrm{L}} = 0 \Leftrightarrow Y(\boldsymbol{x}_i) \subseteq Y_i \tag{8}$$

$$\varepsilon_i^{\mathrm{R}} = 0 \Leftrightarrow Y(\boldsymbol{x}_i) = Y_i \tag{9}$$

すなわち，$Y(\boldsymbol{x}_i) \subseteq Y_i,\ i = 1, 2, \ldots, n$ となる解が存在すれば，第1評価関数が最適値0を取り，$Y(\boldsymbol{x}_i) \subseteq Y_i,\ i = 1, 2, \ldots, n$ の下で，第2評価関数を最小化，すなわち各 $Y(\boldsymbol{x}_i)$ を Y_i に近づける問題になる．

さらに，第2評価関数と問題 (3) の評価関数との関係を調べると，$Y(\boldsymbol{x}_i) \subseteq Y_i,\ i = 1, 2, \ldots, n$ が満足される場合は，ミンコフスキー差モデルの方が必然性モデルより $Y(\boldsymbol{x}_i)$ を Y_i の中央付近に配置する傾向にあることがわかる [2]．また，問題 (7) では，$Y(\boldsymbol{x}_i) \subseteq Y_i,\ i = 1, 2, \ldots, n$ を満足できない場合にも，この条件の違反度が第1評価関数の意味で小さくなる $Y(\boldsymbol{x})$ が得られるというメリットがある．

次に，可能性モデルに対応するものを考えよう．区間の加減算には，元の区間の幅を常に増大するという性質がある．これに注意すると，(4) の区間誤差 E_i^{L} は Y_i の幅が $Y(\boldsymbol{x}_i)$ の幅以上である場合にしか定義できないことがわかる．Y_i の幅が $Y(\boldsymbol{x}_i)$ の幅より小さいときの区間誤差 E_i^{R} は，

$$Y(\boldsymbol{x}_i) = Y_i - E_i^{\mathrm{R}} \tag{10}$$

により定めれば良い．

ミンコフスキー差モデルと同様に，

$$\text{minimize}\sum_{i=1}^{n}|E_i^{\mathrm{R}}| \tag{11}$$

を考え，$|E_i^{\mathrm{R}}| = [\delta_i^{\mathrm{L}}, \delta_i^{\mathrm{R}}]$ として，

$$\text{lex-min}\left(\sum_{i=1}^{n}\delta_i^{\mathrm{L}},\ \sum_{i=1}^{n}\delta_i^{\mathrm{R}}\right) \tag{12}$$

となる辞書式最小化問題を定式化する．このとき，

$$\delta_i^{\mathrm{L}} = 0 \Leftrightarrow Y(\boldsymbol{x}_i) \supseteq Y_i \tag{13}$$

$$\delta_i^{\mathrm{R}} = 0 \Leftrightarrow Y(\boldsymbol{x}_i) = Y_i \tag{14}$$

が得られる．$a_j^{\mathrm{W}},\ j = 1, 2, \ldots, m$ を十分大きくすれば，$Y(\boldsymbol{x}_i) \supseteq Y_i,\ i = 1, 2, \ldots, n$ を満たすので，問題 (12) は $Y(\boldsymbol{x}_i) \supseteq Y_i,\ i = 1, 2, \ldots, n$ の下で，問題 (12) の第2評価関数の最小化，すなわち各 $Y(\boldsymbol{x}_i)$ を Y_i に近づける問題になる．

このように問題 (2) と類似した問題が得られ，やはり線形計画問題に帰着される．これにより得られた $Y(\boldsymbol{x})$ は双対ミンコフスキー差モデルと呼ばれる．第2評価関数と問題 (2) の評価関数との関係を調べると，双対ミンコフスキー差モデルの方が可能性モデルより Y_i が中央付近になるように $Y(\boldsymbol{x}_i)$ を推定する傾向にあることがわかる [2]．

本節で述べた方法を通常データに適用すると，ミンコフスキー差モデルは絶対誤差和最小化によるモデルと一致し，双対ミンコフスキー差モデルは前節で述べた区間回帰モデルと類似したモデルになる．

5．誤差最小化による新しいモデル

式 (4) と式 (10) とを併せると，

$$Y(\boldsymbol{x}_i) + E_i^{\mathrm{L}} = Y_i - E_i^{\mathrm{R}} \tag{15}$$

が得られる．ここで，$E_i^{\mathrm{L}} \neq [0,\ 0]$ であるときは，$E_i^{\mathrm{R}} \neq [0, 0]$ となり，$E_i^{\mathrm{R}} \neq [0, 0]$ であるときは，$E_i^{\mathrm{L}} \neq [0,\ 0]$ となるものと定める．

式 (15) に関しても次の区間誤差和最小化問題を考えることができる．

$$\text{minimize}\sum_{i=1}^{n}|E_i^{\mathrm{L}} + E_i^{\mathrm{R}}| \tag{16}$$

$|E_i^{\mathrm{L}} + E_i^{\mathrm{R}}| = [\gamma_i^{\mathrm{L}}, \gamma_i^{\mathrm{R}}]$ と定め，先と同様に辞書式最小化問題を考えることもできるが，残念ながら，組合せ問題となり，簡単に解くことはできない．代わりに次の問題を考えると，帰着問題は線形計画問題になる．

$$\text{minimize} \sum_{i=1}^{n} \gamma_i^{\mathrm{R}} \tag{17}$$

これにより得られた $Y(\boldsymbol{x})$ は対称ミンコフスキーモデルと呼ばれる．問題 (17) の評価関数に関して，

$$\gamma_i^{\mathrm{R}} = 0 \Leftrightarrow Y(\boldsymbol{x}_i) = Y_i \tag{18}$$

が成立するので，$Y(\boldsymbol{x}_i) = Y_i$，$i = 1, 2, \ldots, n$ をめざしていることがわかる．

式 (4)，(10) では，すべての i に関して共通に，$Y(\boldsymbol{x}_i)$ の幅が Y_i の幅以下，あるいは Y_i の幅以上であることが仮定されたが，式 (15) では，i ごとに，$Y(\boldsymbol{x}_i)$ の幅が Y_i の幅以下であっても，Y_i の幅以上であっても良いことに注意しよう．したがって，問題 (17) は，問題 (7)，(12) より緩く自然な条件の下で，各 $Y(\boldsymbol{x}_i)$ を Y_i に近づける問題と解釈できる．

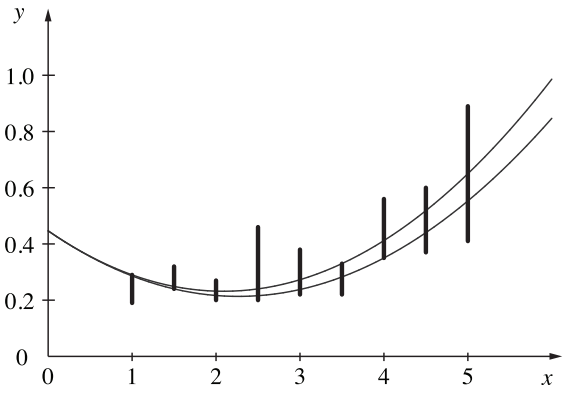
このモデルは通常データに対しても適用でき，絶対誤差和最小化によるモデルと一致する．

6．数値例

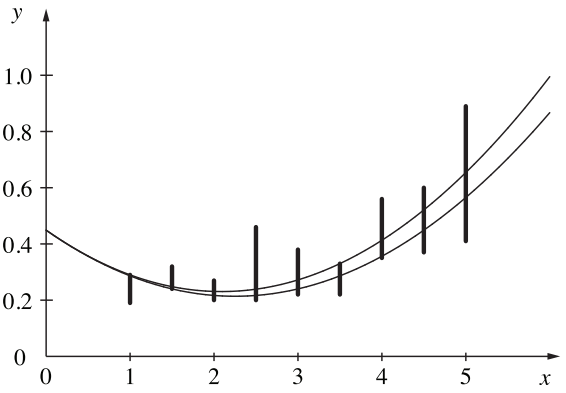
表 1 のデータに対して区間関数 $Y(\boldsymbol{x}) = A_0 + A_1 x + A_2 x^2$ を想定し，各手法で回帰した結果を図 3 に示す．必然性モデルとミンコフスキー差モデル，可能性モデルと双対ミンコフスキー差モデルは，それぞれ同じあるいは類似した結果となる．

表 1: データ

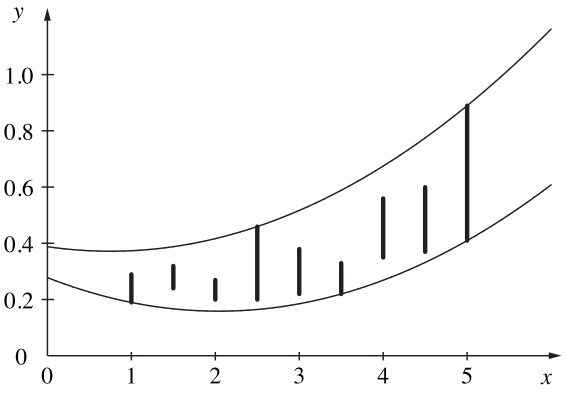
$\boldsymbol{x}_i$	Y_i	$\boldsymbol{x}_i$	Y_i	$\boldsymbol{x}_i$	Y_i
1	[0.19, 0.29]	2.5	[0.2, 0.46]	4	[0.35, 0.56]
1.5	[0.24, 0.32]	3	[0.22, 0.38]	4.5	[0.37, 0.6]
2	[0.2, 0.27]	3.5	[0.22, 0.33]	5	[0.41, 0.89]



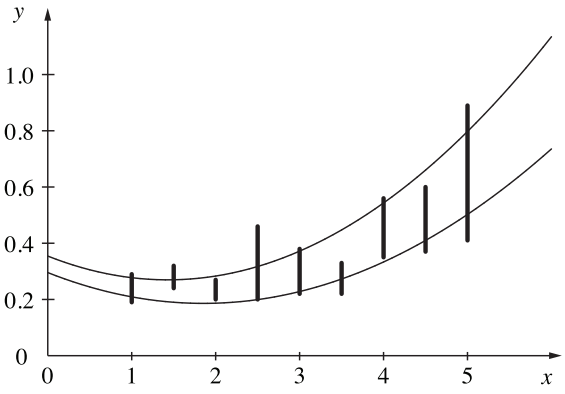
(a) 必然性モデル



(b) ミンコフスキー差モデル



(c) 可能性モデル，双対ミンコフスキー差モデル



(d) 対称ミンコフスキー差モデル

図 3: 区間回帰の結果

7．おわりに

可能性概念を用いた区間回帰分析について述べた．区間回帰モデルでは区間演算の性質から原点から離れた$\boldsymbol{x}_i$ほど，$Y(\boldsymbol{x}_i)$の幅が広くなるという性質があり，原点に近いほど幅が小さかったり，中ほどで幅が最大になる場合には工夫が必要になる．また，異常データを含む場合は，解が得られなかったり，不自然な区間関数が得られてしまう．バラツキをどこまで可能性でとらえるかが課題となる．区間誤差を導入したモデルでは絶対誤差和を用いたが，二乗誤差和やM推定法の評価関数などを導入することもできる．区間回帰分析の考え方は，ニューラルネットワークやSVMにも適用され，数多くの研究がなされている．

本稿では，区間回帰分析を取り上げ，可能性と必然性の概念の適用例を示したが，これらの概念はOR，システム工学，情報工学，知能工学における不確実性を伴う種々の問題に幅広く導入されている．これらの手法は，従来法に対立するものではなく，従来法を補う目的で開発されている．

参考文献

[1] 田中，林，長坂：可能性測度による区間回帰分析，行動計量学，Vol.16, No.1, pp.1－7 (1988).

[2] M. Inuiguchi, T. Tanino: Interval linear regression methods based on Minkowski difference: A bridge between traditional and interval linear regression models, Kybernetika, Vol.42, No.4, pp.423－440 (2006).