No. 945

ファジィ入出力データの可能性線形回帰分析における影響の大きいデータの検出法とその応用

by

Daiki Wakayama and Kazuhisa Takemura

August 2001

ファジィ入出力データの可能性線形回帰分析における影響の大きいデータの検出法とその応用[1]

若山 大樹[2]
(筑波大学大学院 社会工学研究科)

竹村 和久[3]
(筑波大学 社会工学系)

Abstract :

We present a detection method of influential observations for possibilistic linear regression analysis, where input data and output data are represented by L-R fuzzy numbers. In this method, we use a sensitivity analysis using concepts of duality and its allowable right-hand side range in linear programming, in order to evaluate an effect of influential observations on fuzzy regression coefficients. The proposed method deals with two constraints corresponding to one observation using the dual variable and their allowable right-hand side ranges in primal problem. We demonstrate an application of the proposed method on consumer decision research that examines effects of multi-attributes on behavioral intention.

Key word :

Influential observations, Possibilistic linear Regression Analysis, Group of Inliers,

---

[1] A Detection Method of Influential Observations for Possibilistic Regression Analysis for Fuzzy Input-Output Data and Its Application

[2] Wakayama Daiki (Doctoral Program in Policy and Planning Sciences, University of Tsukuba)

〒305-8573 つくば市天王台 1-1-1 筑波大学大学院 社会工学研究科

Doctoral Program in Policy and Planning Sciences, University of Tsukuba

1-1-1 Tennodai Tsukuba Ibaraki 305-8573, Japan

[3] Takemura Kazuhisa (Institute of Policy and Planning Sciences, University of Tsukuba)

〒305-8573 つくば市天王台 1-1-1 筑波大学 社会工学系

Institute of Policy and Planning Sciences, University of Tsukuba

1-1-1 Tennodai Tsukuba Ibaraki 305-8573, Japan

要約：
入出力データがファジィ数である場合の可能性線形回帰分析における影響の大きいデータの検出法を提案する。提案した方法では、一つの観測値がファジィ回帰係数に与える影響の大きさを評価するために、線形計画問題における双対変数と許容幅の概念を利用し、一つの観測値に対応する二本の制約式をファジィ数演算にもとづいて同時に取り扱う方法である。提案した方法を用いて、行動意図に与える多属性態度の影響を調べた消費者調査データへの応用例を示す。
キーワード：　影響の大きいデータ、可能性線形回帰分析、データ群内部データ

1．はじめに

可能性線形回帰分析は、入出力関係をあらわすシステム構造自体に曖昧性があると考えられる場合に用いられる分析手法[13,14,19]であり、データセットには複数のデータ群ではなく一つのデータ群を仮定する。可能性線形回帰分析の応用例として、ある消費者集団の持つ評価構造を可能性線形システムで表現した例があげられる[18]。一般に、得られた集計データ[10]は、同質性を仮定することであたかも一人の個人が繰り返し行った結果であるかのように取り扱って分析することができる。しかしながら、実データには、仮定した同質的な評価構造を持つ被験者集団の中に異質な評価構造を持つ被験者のデータが含まれることがある。異質なデータは、可能性線形回帰分析において、たとえ少数であっても、結果に大きな影響を与えることがあるため[6,12,22]、消費者行動のような集計データを取り扱う分野においては、常に影響の大きいデータの存在を考慮する必要がある。

本論文では、集計データに含まれる、回帰係数に与える影響の大きいデータを検出するための方法を提案し、消費者の多属性調査データへの応用例を提示する。

一般に、線形計画法に基づく可能性線形回帰分析は、一つの観測値が二本の制約式を構成しており[13,14,19]、一つの観測値が回帰係数に与える影響を調べるためには、二本の制約式を同時に取り扱う必要がある。ところが、線形計画法の感度分析によって得られる制約式右辺に関する許容幅は、一本の制約式のみ取り扱うことが前提であるから[15]、二本の制約式を同時に取り扱うことは、双対変数や許容幅の性質が損なわれる可能性がある。そこで、本論文では、一つの観測値が構成する二本の制約式をファジィ数演算にもとづいて同時に取り扱っても双対変数や許容幅の性質が損なわれない方法により、回帰係数に大きな影響を与えるデータの検出方法を提案する。

本論文では、まず次の2章において、可能性線形回帰分析における回帰係数に与える影響の大きいデータについての取り扱いに関する先行研究とその問題点について述べる。次に、3章において、坂和･矢野の可能性線形回帰分析[14]の概説を行い、渡邊･小沢の提案した説明力の指標[20]をもちいてデータ数を考慮した説明力の指標を提案する。4章では、双対変数と許容幅を用いた特異データの検出法を提案し、最後の5章では、数値例と消費者の多属性調査データへの応用例を示す。

2．データとモデルへの影響

可能性線形回帰モデルは確率モデルではないので、観測誤差の概念はなく、たとえ外れ値や異常値のような大きな誤差を持つデータが存在したとしても、入出力関係を表すシステムの可能性とみなされて処理される。その結果、すべての入出力関係が実現可能なように推定された回帰係数(ファジィ係数)が広がることで、モデルの説明力[20]は大きく低下する。

可能性線形回帰モデルを構成するにあたり、薮内らは観測データを特異データ、周辺データ、データ群内部データの三つに分類している[21]。ここで特異データとは、考慮している(同質的あるいは単一の)システムの構造とは異なる原因で生成されたと考えられるデータの総称[21,22]であり、回帰係数に大きな影響を与えるデータのことである。周辺データは可能性線形回帰モデルを規定しているデータであり、これを取り除いたとき回帰係数に影響を与えるデータである。他方、データ群内部データはモデルを規定していないデータであり、これを取り除いても回帰係数に影響を与えることはない[21,22]。周辺データとデータ群内部データは、各データの構成する制約式の最適解における等号関係で分類することができる。したがって、周辺データを取り扱うことで影響の大きい

データが回帰係数に及ぼす影響を効果的に取り扱うことができる[22]。また、パラメータの符号に関する知識を制約条件に反映させて、影響の大きさを評価することができる点においても、従来の統計手法[1,3]に依らない検出方法として、可能性回帰分析独自に進展してきた。

可能性線形回帰分析において、回帰係数に大きな影響を与えるデータの取り扱いに関する手法は、ロバスト推定的方法[9,21,22]と回帰診断的方法[5,8]に大別できる。これらは、いずれも外れ値の処理に関する従来の統計手法[1,3]ではなく、可能性線形回帰分析におけるパラメータ推定のための線形計画法による定式化を基礎に展開したものである。Lee&Tanaka[8]は、外れ値の検出において双対変数を用いることの有効性を指摘している。しかしながら、回帰係数に与える影響の大きいデータを検出するための方法としては双対変数だけではなく、感度分析における許容幅の両方の概念を用いる必要がある。また、双対変数と許容幅の両方の概念を用いて分析する際には、回帰係数に影響を与えるデータの中から、特異データ(外れ値)がいくつ含まれるのかについての客観的な決め方が必要である。このためには、可能性線形回帰分析の説明力の指標についての考え[20]を取り入れることが必要であると考えられる。

3．可能性線形回帰分析と説明力の指標

3．1 入出力ファジィデータの可能性線形回帰分析

坂和･矢野[14]に基づき、可能性線形回帰分析について概説を行うと以下のようになる。

i 番目の入出力ファジィデータセット(被説明変数と説明変数の組)を次のように定義する。

$$\left(Y_i; X_{i0}, X_{i1}, X_{i2}, \ldots, X_{im}\right),\ i=1,2,\cdots,n \tag{1}$$

ここで、$X_{i0}=1$、ファジィ係数を A_j、$\left(j=0,1,\cdots,m\right)$とすると、可能性線形回帰式は、次式のように表される。

$$Y = A_0 + A_1 \otimes X_1 + \ldots + A_m \otimes X_m \tag{2}$$

ここで、$\otimes$ は2つのファジィ数 A_j と X_j の拡張原理に基づく積の演算を表している。$z^L_{i(\alpha)}$ を予測区間の下限、$z^R_{i(\alpha)}$ を予測区間の上限とすると、α レベルにおける予測値 Z_i は、閉区間、

$$Z_i = \left[z^L_{i(\alpha)}, z^R_{i(\alpha)}\right],\quad \alpha \in (0,1] \tag{3}$$

で表される。ただし、

$$z^L_{i(\alpha)} = \sum_{j=0}^{m}\left\{\min\left(a^L_{j(\alpha)}x^L_{ij(\alpha)}, a^L_{j(\alpha)}x^R_{ij(\alpha)}\right)\right\} \tag{4}$$

$$z^R_{i(\alpha)} = \sum_{j=0}^{m}\left\{\max\left(a^R_{j(\alpha)}x^L_{ij(\alpha)}, a^R_{j(\alpha)}x^R_{ij(\alpha)}\right)\right\} \tag{5}$$

ここで、$a^L_{j(\alpha)}x^L_{ij(\alpha)}$ は第 j 番目の属性におけるファジィ係数の α レベルにおける下限(最小値)と第 i, j 番目の入力データの α レベルにおける下限(最小値)との積を表している。

ここで、予測値と観測値(出力データ)の適合度が、

$$Pos\left(Z_i = Y_i\right) \geq h \tag{6}$$

のもとで、予測値の幅の総和の最小化問題(以下、$FLP\left(h; J_1, J_2, J_3\right)$ と呼ぶ)を解き、ファジィ係数の推定値を求める。

$FLP\left(h; J_1, J_2, J_3\right)$ の目的関数：

$$Min \quad \sum_{i=1}^{n}\left\{\left(z^L_{i(\alpha=1)} - z^L_{i(\alpha=0)}\right) + \left(z^R_{i(\alpha=0)} - z^R_{i(\alpha=1)}\right)\right\} \tag{7}$$

制約条件：

$$-z^L{}_{i(h)} \geq -y^R_{i(h)} \tag{8}$$

$$z^R{}_{i(h)} \geq y^L_{i(h)} \tag{9}$$

$$a^L_{j(h)} \geq 0,\ j \in J_1 \tag{10}$$

$$a^L_{j(h)} \leq 0,\ a^R_{j(h)} \geq 0,\ j \in J_2 \tag{11}$$

$$a^R_{j(h)} \leq 0,\ j \in J_3 \tag{12}$$

$$-a^L_{j(h)} + a^R_{j(h)} \geq 0 \tag{13}$$

$$j \in \{0,\cdots,m\} = J_1 \cup J_2 \cup J_3,$$
$$J_1 \cap J_2 = \phi, \quad J_2 \cap J_3 = \phi, \quad J_3 \cap J_1 = \phi, \tag{14}$$

ただし、$i = 1,\ldots,n$。ここで、パラメータが左右非対称の三角ファジィ数と仮定するとき、式(7)から式(13)に、以下の式(15)、式(16)を導入して線形計画問題を解く(左右対称のときは、$c_j^L = c_j^R$)。

$$a_{j(h)}^L = a_j^M - (1-h)c_j^L \tag{15}$$

$$a_{j(h)}^R = a_j^M + (1-h)c_j^R \tag{16}$$

ただし、a_j^M、c_j^L、c_j^Rは、それぞれ、三角ファジィ数の代表値(左右対称の場合は中心)、代表値から下限値までの幅、代表値から上限値までの幅を表している。

ここで、i番目のデータは、二つの制約式を構成し、式(8)が予測値の下限、式(9)が予測値の上限を表している。式(10)から式(14)のJ_1, J_2, J_3は、属性番号の集合を表しており、式(10)は係数の符号が非負の属性、式(11)は係数の符号の上限が非負で下限が非正の属性、式(12)は係数の符号が非正の属性を表している。式(13)は、係数の上限値が下限値よりも小さくならないということであり、式(14)はJ_1, J_2, J_3がたがいに共通要素を持たず、属性番号$\{0,\cdots,m\}$は集合J_1, J_2, J_3のいずれかに必ず属しているということを意味している。

パラメータの正・負の符号が事前にわかっている場合、それに対応する添え字の集合J_1, J_2, J_3を設定するが、事前知識がない場合には、対応する添え字集合のすべての組み合わせに対して、線形計画問題$FLP(h; J_1, J_2, J_3)$を解き、それらの中で目的関数が最も小さくなる最適解の添え字集合の組み合わせに決める[13,14]。

ここで、スラック変数$s_i^{(L)}, s_i^{(R)}$を導入し、最適解における制約式の不等号を等号に直すと、

$$-\hat{z}_{i\alpha}^L - s_i^{(L)} = -y_{i\alpha}^R \tag{17}$$

$$\hat{z}_{i\alpha}^R - s_i^{(R)} = y_{i\alpha}^L \tag{18}$$

$\hat{z}_{i^*\alpha}^L$は最適解における予測値の下限、$\hat{z}_{i^*\alpha}^R$は最適解における予測値の上限とする。

3．2 説明力の指標

本論文では、渡邊らの提案した説明力の指標[21]を用いる。渡邊らに従い説明力をR^2で表すと、

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(z_i^R - z_i^L\right)}{n(\max y - \min y)}, \quad y \in \bigcup_1^n Y_i \tag{19}$$

ここで、$\max y, \min y$はそれぞれ出力データの最大値、最小値を表している。ここで、出力データがファジィ数のとき、渡邊らの提案した説明力の指標を出力データがファジィ数である場合に拡張すると、式(20)のように定義される。

$$R_\alpha^2 = 1 - \frac{\sum_{i=1}^{n}\left(z_{i\alpha}^R - z_{i\alpha}^L\right)}{n\left(\max y_\alpha^R - \min y_\alpha^L\right)}, \quad \left[y_\alpha^L, y_\alpha^R\right] \subset \bigcup_1^n Y_i \tag{20}$$

ここで、$\max y_\alpha^R, \min y_\alpha^L$は出力データの$\alpha$レベル集合に含まれる要素の上限値の最大値、下限値の最小値をそれぞれ表している。ここで、データ数を考慮した説明力の指標を次のように定義する。

$$R_\alpha^2(n,p) = \frac{(n-p)}{n} R_{\alpha,p}^2, \quad p = 0,1,\cdots \tag{21}$$

ただし、$R_{\alpha,p}^2$は、特異データp個のときの説明力、$n-p$は処理後のデータ数を表す。この基準は、特異データの個数を検討するために用いる。

4．影響の大きいデータの評価

回帰係数に与える影響の大きいデータを検出する方法について以下にその概略を述べる。

1. 主問題$FLP(h; J_1, J_2, J_3)$と双対問題を解き、主問題のスラック変数$s_i^{(L)}, s_i^{(R)}$が零(ゼロ)の制約

式に対応する双対変数と、制約式右辺の制約式右辺に関する感度分析により許容幅(制約式右辺の値を減少させる方向の幅：許容減少幅)を求める。
2. 特異データの候補(周辺データ)を求める。
3. 影響の大きさを双対変数と許容幅の積で評価し、影響最大の周辺データを抽出する。(なお、最適解におけるスラック変数、双対変数と許容幅は線形計画問題を解くことが可能な市販のソフトウエアで出力結果を得ることができる。これらの定義や求めるための公式については付録1を参照されたい)

制約式(8)(右辺は y_α^R)に対応する許容幅(制約式右辺の値を減少させる方向の幅：許容減少幅)を $\Delta y_i^{(R)}$、制約式(9)(右辺は y_α^L)に対応する許容幅を $\Delta y_i^{(L)}$ で表し、双対変数を $w_i^{(R)}, w_i^{(L)}$ で表す。ここで、括弧のついた $(R),(L)$ に、上限や下限という意味はなく、括弧のないものと区別する。

以下に回帰係数に与える影響の大きいデータを抽出するための手続きを述べる。

t $(t=0,1,2,\cdots)$ を計算回数とし、データ番号を $i \in I$ とする。ここで、計算回数tにおける周辺データ[21]の番号の集合を I_t^* とする。

Step0： $FLP(h;J_1,J_2,J_3)_t$ を解く。

Step1： $FLP(h;J_1,J_2,J_3)_t$ のスラック変数がゼロ(零)の制約式を持つデータ番号を求める。計算回数tにおける周辺データ番号の集合 $i^* \in I_t^*$ は、

$$i^* = \left\{ i \,\middle|\, \left(s_i^{(L)}=0\right) or \left(s_i^{(R)}=0\right), i \in I \right\} \quad (22)$$

Step2： 計算回数0からtにおける周辺データ番号の集合 $(I_0^*,\cdots,I_{t-1}^*,I_t^*)$ の共通集合が、

$$I_0^* \cap \cdots \cap I_{t-1}^* \cap I_t^* \neq \phi \quad (23)$$

のとき、Step3へ。

$$I_0^* \cap \cdots \cap I_{t-1}^* \cap I_t^* = \phi \quad (24)$$

のとき、回帰係数への影響度最大のデータを求め終了。ただし、ϕ は空を表している。

Step 3： i^* に該当する制約式の双対変数と許容幅をもとめ、これらの積が最大の制約式を構成するデータ番号を求める。

$$i_{\max}^* = \left\{ i^* \,\middle|\, Max\left(w_{i^*}^{(R)} \times \Delta y_{i^*}^{(R)}, w_{i^*}^{(L)} \times \Delta y_{i^*}^{(L)}\right), \; i \in \left(I_0^* \cap \cdots \cap I_{t-1}^* \cap I_t^*\right) \right\} \quad (25)$$

Step4：周辺データ $i_{\max}^*$ がモデルの下限を規定している場合と上限を規定している場合について場合わけする。

$\left(w_{i_{\max}^*}^{(R)} \times \Delta y_{i_{\max}^*}^{(R)}\right) \neq 0, \left(w_{i_{\max}^*}^{(L)} \times \Delta y_{i_{\max}^*}^{(L)}\right) = 0$ のとき、Case1へ。

$\left(w_{i_{\max}^*}^{(R)} \times \Delta y_{i_{\max}^*}^{(R)}\right) = 0, \left(w_{i_{\max}^*}^{(L)} \times \Delta y_{i_{\max}^*}^{(L)}\right) \neq 0$ のとき、Case2へ。

Case1：ファジィ数(観測データ) $Y_{i_{\max}^*}$ を $Y'_{i_{\max}^*}$ に置き換える。

$$Y'_{i_{\max}^*} = Y_{i_{\max}^*} + \left(\Delta y_{i_{\max}^*}^{(R)} + \varepsilon\right) \quad (26)$$

したがって、制約式右辺はファジィ数の演算より、

$$-y'^R_{i_{\max}^*(h)} = -y^R_{i_{\max}^*(h)} - \left(\Delta y_{i_{\max}^*}^{(R)} + \varepsilon\right), \quad (27)$$

$$y'^L_{i_{\max}^*(h)} = y^L_{i_{\max}^*(h)} + \left(\Delta y_{i_{\max}^*}^{(R)} + \varepsilon\right) \quad (28)$$

Case2：ファジィ数(観測データ) $Y_{i_{\max}^*}$ を $Y'_{i_{\max}^*}$ に置き換える。

$$Y'_{i_{\max}^*} = Y_{i_{\max}^*} - \left(\Delta y_{i_{\max}^*}^{(L)} + \varepsilon\right) \quad (29)$$

すなわち、制約式右辺はファジィ数の演算より、

$$-y'^R_{i_{\max}^*(h)} = -y^R_{i_{\max}^*(h)} + \left(\Delta y_{i_{\max}^*}^{(L)} + \varepsilon\right), \quad (30)$$

$$y'^L_{i_{\max}^*(h)} = y^L_{i_{\max}^*(h)} - \left(\Delta y_{i_{\max}^*}^{(L)} + \varepsilon\right) \quad (31)$$

ただし、$\varepsilon\,(\varepsilon > 0)$ は十分に小さな数。

Step 5： t を $t+1$ に置き換えてstep0へすすむ。

Step2では、計算回数が増えるに従って共通集合に含まれる要素は単調に減少し、有限回で終了する。Step3では、スラック変数がゼロ(零)の制約式に該当する双対変数と許容幅を用いて、回帰

係数に与える影響の大きさを式(25)により評価する。双対変数と許容幅がともに非ゼロである制約式は相補性定理よりスラック変数が必ずゼロになり、この制約式右辺を変化させたときは必ず回帰係数が変化する。回帰係数に与える影響の大きさは、双対変数と許容幅の積の大きさに加えて、説明力の変化の大きさを検討する。Step4 で、許容幅よりも ε (十分に小さな数)だけ余分に変化させる理由は、次の Step1 で、スラック変数がゼロになるのを避け、周辺データとデータ群内部データの区別を容易にするためである。双対変数と許容幅の積の大きさは、周辺データとデータ群内部データの隔たり量を近似的に表している。

ところで、Step4 において、双対変数と許容幅の積が最大の周辺データ番号の構成する二本の制約式を同時に変化させており、一見したところ許容幅を用いることに問題点があるように見える。なぜなら、双対変数と許容幅は、一つの制約式以外の他の制約式は変化させないことが前提の概念であるからである。しかしながら、Step4 が正当な手続きであるということを証明するのは困難ではない。この証明は付録2に示す。

5．応用例

5．1　数値例

以下の数値例は、Lee&Tanaka の数値例[8]をもとに 13 番目のデータを一つ加えて作成したものであり、影響の大きいデータを評価するために、双対変数の大きさだけでなく許容幅の概念も用いる必要があることを示すためのものである。

表1. 数値例

ID	1	2	3	4	5	6	7	8	9	10	11	12	13
X	1	2	3	4	5	6	7	8	9	10	11	12	7.5
Y	7	8	8	12	9	10	7	10	11	12	12	13	7.5

最適解における双対変数の値は、大きいものから順に、データ番号 $i=7$ 番(12.1)、4 番(8.8)、12 番(4.2)、1 番(0.92)となる。ただしデータ番号横のカッコ内は、双対変数の値である。ここで、Lee&Tanaka の方法[8]では、最も大きい双対変数の値を持つデータ番号が選ばれるから 7 番目のデータが選ばれる。

ところが、ひとつのデータを取り除いたときの説明力[20]の変化は、7 番目のデータを取り除いたとき、11.3 から 18.5 になるが、4 番目のデータを取り除いたときには、11.3 から 43.4 になり、7 番目を取り除いたときよりも大きく改善される。したがって、一つのデータを取り除いたときの影響の大きさは 4 番目のデータが最も大きく、双対変数の大きさだけで影響の大きさを評価するのは問題があることがわかる。

許容幅の概念を用いた本手法によると、Step1 で周辺データ番号(1,4,7,12)が選ばれ、各周辺データについて双対変数と許容幅の積を計算(Step3)し、値の大きいものからデータの構成する制約式右辺を変化させる(Step4)。計算回数 4 回目で終了し、本手法による評価値の大きいものから順に、データ番号 $i=$ 4 番(26.4)、7 番(5.6)、1 番(0.12)、12 番(0.0)の順になった。ただしデータ番号横のカッコ内は、双対変数と許容幅の積である。これは、ひとつの周辺データを取り除いたときの説明力の変化の大きさの順に一致している。したがって、影響の大きいデータを評価するためには、双対変数の大きさだけでなく許容幅の概念も用いる必要があるといえる。ここで、12 番のように評価値が 0.0 となっているデータは、それ自身を変化させることなく Step2 において共通集合の要素である周辺データ番号から外れたデータであり、他の影響の大きいデータに依存して見かけの上で周辺データになっていたことを意味している。

5．2　集計データへの適用

調査データは、衣服購入の際の地域選択に関する意識調査と題して、私立女子大生 44 人に対して、購買地域の多属性に関する評価が総合的な評価(購買意図)とどのような関係にあるのかを明らかにすることを目的として得られたものである。予備調査をもとに、衣服購入の際の地域や属性について決定し、質問紙が作成された[16,17]。被験者に、①衣服購入の際の対象地域に関する 8 個の属性項目(価格、品揃え、交通の便、お洒落さ、上品さ、新しさ、落ち着き、華やかさ)について評価させ、②各購買地域がそれらの属性をどの程度有しているかを評価させ、③衣服購入の際に各地域に行くことについての評価と行動意図を評価させたものである。評価の方法は、評価における曖昧さを測定可能なように考案されたファジィ評定法(ファジィグラフ評定尺度図法)である[4,17]。ファジィ評定法によって得られたデータは、ファジィ評定データと呼ばれている。竹村は、これらの調査データに対して被験者ごとに、線型加算型の多属性態度モデル[7]にもとづく個人単位分析を行った[16,17]。本論文では、個人単位分析結果を事前知識として用いて、定数項以外は正の符号を仮

定した線形加算型モデルにより、被験者(消費者)グループについての多属性評価モデル(可能性線形システム)を同定するための分析を行う。まず、対象地域についての被験者の購買意図評価をシステムの出力としての総合評価、8 属性に関する評価をシステムの入力と仮定する。次に、同質的な総合評価システムを持つ被験者集団から得られたデータであると仮定する。

ファジィ多属性態度モデルは、通常の多属性態度モデルをファジィ評定データで分析できるように拡張したものである。構成された態度尺度によって購買意図(ここでは総合評価)Y を説明するために、対象に対する j 番目の属性評価を X_j、対象に対する評価に寄与する重みを W_j とすると、

$$Y = W_1 \otimes X_1 + W_2 \otimes X_2 + \cdots \qquad (32)$$

で表される加算型線形モデルである。この多属性態度モデルは、すべての被験者が同質的に答えると仮定した個人単位の分析である[7]。ここで、多属性態度モデルにおいて観測されたデータは、Y、X_j、W_j であり、これに対して、集計データに同質的なシステムを仮定し可能性線形回帰分析を行う場合、観測変数は Y、X_j である。集計データの同質性を代表するような W_j (入出力システムの関係)を集計データに含まれる影響の大きいデータの影響を考慮して推定することが本適用例の目的である。

5．2．1 特異データの検討

①個人単位分析の結果 72.7%の被験者に当てはまったとの報告[16,17]がある。②ファジィ回帰係数の結果、回帰係数の(下限,上限)が属性 1 から順に(0,0)、(0,0.063)、(0,0)、(0,0.37)、(0,0)、(0,0.01)、(0,0)、(0,0)となり、すべての属性の下限がゼロとなった。ただし、適合度基準 $h = 0$ を用いた。③説明力[19]は、27.12%であった。この説明力は、同じ線形モデルでかつ個人別分析に用いた情報量が少ないことを考慮しても、72.7%からあまりにかけ離れている。以上の三つの結果から、同質性の仮定のもとで集計データに特異データや影響の大きいデータが含まれる可能性について調べる必要がある。

5．2．2 適用結果

モデルに与える影響の最も大きなデータを一つづつ順に取り除いたときの説明力の変化と処理後のデータ数が全データ数に占める割合、およびデータ数を考慮した説明力は、表2のようになる。ここで、特異データの個数がゼロのときの説明力は 27.1 である。同様に、特異データの個数が 1 個のとき、説明力は 38.7 に改善し、特異データ処理後のデータ数が全データに占めるの割合は 0.98(=43/44)、データ数を考慮した説明力は 37.8 になる。図1、図2は表2の特異データ数と説明力およびデータ数を考慮した説明力をグラフにしたものである。ここで、データ数を考慮した説明力は、最大となる特異データの数は 9 個であるが、図2より特異データが 6 個から 9 個まではあまり変化がない。したがって、特異データ数を 6 個とし、このときの回帰係数を表3に示す。

表2. 特異データの個数と説明力、

特異データ数 p	0	1	2	3	4
説明力 $R^2_{n,p}$	27.1	38.7	46.7	50.1	57.4
処理後のデータ数の割合 $(n-p)/n$	1.0	0.977	0.955	0.932	0.909
データ数を考慮した説明力 $R^2_{n,p}(n,p)$	27.1	37.8	44.6	46.7	52.2

特異データ数 p	5	6	7	8	9	10
説明力 $R^2_{n,p}$	64.0	71.0	74.4	76.7	79.4	80.6
処理後のデータ数の割合 $(n-p)/n$	0.886	0.864	0.841	0.818	0.795	0.773
データ数を考慮した説明力 $R^2_{n,p}(n,p)$	56.7	61.3	62.5	62.7	63.2	62.3

表3. 特異データの個数と回帰係数

特異データ数	0	0	0	6	6	6
	下限	代表値	上限	下限	代表値	上限
定数項	0	0	0	0.23	2.2	4.17
価格	0	0	0	0.13	0.13	0.13
品揃え	0	0.31	0.63	0	0	0.01
交通の便	0	0	0	0.08	0.09	0.1
お洒落さ	0	0.18	0.37	0.42	0.42	0.42
上品さ	0	0	0	0	0	0
新しさ	0	0	0.01	0	0.08	0.16
落ち着き	0	0	0	0	0	0
華やかさ	0	0	0	0	0	0

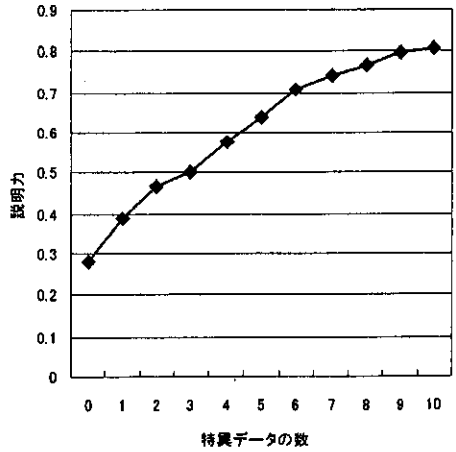

図1：特異データの個数と説明力(%)

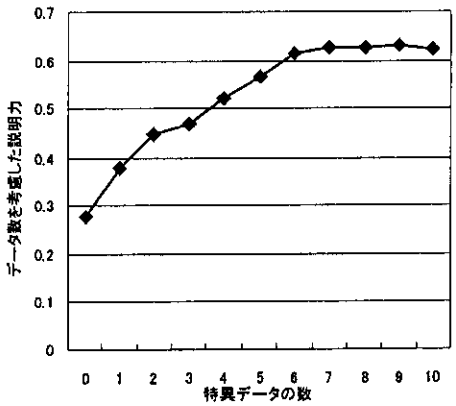

図2：データ数を考慮した説明力(%)

5．2．3 可能性線形回帰分析における特異データ検討の意味と解釈のための仮定

集計データに含まれる特異データは、同質的なシステムにより生成された可能性が全くないとは言い切れない。本来ならば引き続きデータを収集して判断する必要があるが、現時点で解釈しなければならない場合、影響の大きいデータを特異データと解釈するためには、以下のような理由により、解釈のための仮定を置く必要がある。第一の理由は、仮に複数のデータ群が存在しているならば、そもそもこの集計データに同質性を仮定することはできないので可能性線形回帰モデルを構成することはできない。これは、集計データの背後に質の異なる複数の混合分布が存在することを認めることになるからである。第二の理由は、特異データが仮にすその重い単峰型の確率分布により出現したと考えられる場合、引き続きデータを収集したときに、データがある頻度でデータ群と特異データの間に出現する可能性があるからである。以上のことから、特異データの処理は、回帰係数に与える影響の大きさだけで判断するのではなく、生成された背後に混合分布やすその重い分布によらず出現したと仮定しなければならない。提案した手法では、双対解と許容幅の積を用いているため、周辺データがデータ群を成している場合、影響の大きいデータとして検出されにくいという特徴がある。図1から、処理した影響の大きいデータの個数にしたがって単調に説明力が増加していることがわかる。ここで、取り除いた特異データ数を考慮に入れた説明力(3.2 節の式(21)で定義した)は、6個から9個まで大きな変化はみられないが最大値9個を超えると減少に転じている(図2)。このようにデータ数を考慮した説明力が最大値を持つことは、説明力が定義上100パーセントを超えることはないこと、データ数を考慮すると特異データ数が増えるに従って分析に用いられるデータ数が減少することからも理解できる。

この適用例では、三つの先験情報を出発点とし、可能性線形回帰モデルを構成する際に特異データが含まれる可能性について調べ、回帰係数に与える影響の大きいデータの検出と説明力や回帰係数の変化、特異データの個数のきめ方と解釈のために必要な仮定を述べた。この適用例で用いた特異データの個数の決め方には、仮に目的関数値や説明力に極端な変化が見られれば、石渕・田中の方法[5]が有効であろう。しかしながら、この適用例のように、含まれる特異データの個数について、必ずしも境界が鮮明ではない可能性があるので、一つの目安としてデータ数を考慮した説明力を判断の基準に用いている。

6．まとめ

本論文では、調査データに特異データが含まれる可能性のあるときの、可能性線形回帰分析における特異データの発見と処理に関する判断支援のための簡便法を提案し、消費者行動分析における問題解決のための方法として集計データへの適用例を示した。集計データに対して同質性を仮定できるなら、可能性線形回帰分析を用いることができる。提案した方法は、回帰係数に大きな影響を

与えるデータを双対変数や許容幅の概念を用いて評価する回帰診断的方法である。

回帰係数に与える影響の大きいデータを調べるためには、双対変数の値の大きさだけでなく許容幅の概念を用いる必要があることについて数値例を用いて述べ、先行研究で十分に扱えない影響力の大きさを加味した分析ができることが示された。一方で、集計データに対して仮定した同質性の仮定がどの程度満たされるかを議論するため、可能性線形回帰分析における特異データの検討の意味と解釈のための仮定について述べた。4章および付録では、周辺データとデータ群内部データの分類と双対変数や許容幅との間にある理論的関連性について述べた。1つのデータには予測値の上限と下限を表す二つの制約式があり、1つの周辺データの出力を変化させて回帰係数への影響をしらべるためには、二つの制約式を同時にファジィ数の演算に基づいて変化させる必要がある。1つの制約式右辺だけを変化させることが前提の許容幅の概念が、二つの制約式を同時に変化させても許容幅の性質を失うことなく有効であることを示し、回帰係数に与える影響の大きいデータの検出法を提案した。

本論文で提案した方法は、回帰係数に与える影響の大きいデータを検出するためのものであるが、影響の大きい特異データが含まれることで、データ群内部データが周辺データとして検出される問題や、データ群内部データが周辺データや特異データとして検出されてしまう可能性を検討する必要がある。以上の点は今後の課題である。

[謝辞]
本論文を作成するにあたり、筑波大学宮本定明先生、門田安弘先生、松尾博文先生には大変有益なコメントをいただきました。記して感謝申し上げます。

[参考文献]

[1] Belsley, D. A., Kuh, E., Welsch, R. E.： Regression Diagnostics. － Identifying influential data and sources of collinearity －, New York: John Wiley&Sons, (1980).

[2] Hartley, R.V ： Operations Research, Goodyear Publishing Company, Inc.,(1976). [門田安弘, 加登豊共訳,数理計画法の経営活用, 新東洋出版社, pp.147-159. (1979)]

[3] Hawkins D. M.： Identification of outliers. Chapman and Hall,(1980)

[4] Hesketh, B., Pryor, R., Gleitzman, M., and Hesketh, T.： Practical applications and psychometric evaluation of a computerised fuzzy graphic rating scale. In T.zetenyi(Ed.), Fuzzy Sets in Psychology. New York: North Holland, pp.425-424, (1988).

[5] 石渕久生,田中英夫： 混合 0-1 整数計画問題による区間回帰分析,日本経営工学会誌, Vol.40,No.5,pp.312-319,(1989).

[6] Kim, K., Moskowitz, H., and Koksalan, M： Fuzzy versus statistical linear regression, European Journal of the Operations Research, 92, pp.417-434, (1996).

[7] 小島健司：多属性態度と行動意図モデル, 中西正雄(編著),消費者行動分析のニューフロンティア(第2章), 誠文堂新光社,(1984).

[8] Lee, H. and Tanaka, H.：Dealing with outliers by fuzzy regression reflecting central tendency 第13回ファジィシステムシンポジウム予稿集, pp.365-366, (1997).

[9] Lee, H., and Tanaka, H.： Upper and lower approximation models in interval regression using regression quantile techniques, European Journal of the Operations Research, 116, pp.653-666, (1999)

[10] 中西正雄：多属性分析, 中西正雄(編著),消費者行動分析のニューフロンティア(第6章), 誠文堂新光社, (1984).

[11] 西田俊夫：ORハンドブック,pp.144-146, 朝倉書店,(1972).

[12] Peters, G. ： Fuzzy linear regression with fuzzy intervals, Fuzzy Sets and Systems, 63,pp.45-55, (1994).

[13] 坂和正敏：ファジィ理論の基礎と応用（5章・7章）, 森北出版, (1989)

[14] Sakawa, M. and Yano,H.：Fuzzy linear regression analysis for fuzzy input-output data, Information Sciences, 63, pp.191-206, (1992)

[15] Schrage, L. ： LINDO(4th), Scientific Press ,(1991) [新村秀一・高森寛訳： 実践数理計画法, pp33-44,朝倉書店,(1992)]

[16] 竹村和久：ファジィ多属性態度モデルによる

購買地選択の分析について ― エリア・マーケティングのための消費者心理測定の提案―, 地域学研究, Vol.22, pp.119-132, (1992).
[17]竹村和久：意思決定の心理, 福村出版, (1996).
[18]Takemura.K ：A fuzzy linear regression analysis for fuzzy input-output data using the least squares method under linear constraints and its application to fuzzy rating data, Journal of Advanced Computational Intelligence, 3(1), (1999).
[19]田中英夫：ファジィ回帰モデル, 寺野, 菅野, 浅居(共編)：ファジィシステム入門, 第 4 章 pp.67-81, オーム社,(1987)
[20]渡邊匠,小沢和浩：可能性回帰分析における評価指標の定義, 日本ファジィ学会誌, Vol.11,No.1,pp99-103,(1999).
[21]薮内賢之,和多田淳三：超楕円関数に基づくファジィロバスト回帰分析, Journal of the Operations Research Society of Japan, Vol.39, No.4, (1996).
[22]薮内賢之,和多田淳三,辰巳憲一:誤差データのファジィ回帰分析, 日本ファジィ学会誌, Vol.6, No.6, (1994).

付録1 双対変数と許容幅の定義

双対変数は、主問題$FLP(h;J_1,J_2,J_3)$に対する双対問題を解くことにより得られる。双対問題は以下の定式化により得られる。

$$Max:\ \sum_i^n\left(-y_{i(h)}^R w_i^{(R)}+y_{i(h)}^L w_i^{(L)}\right) \quad (33)$$

Subject to:

$$-\sum_i^n x_{ji\alpha}^{(L)} w_i^{(R)}-w_j\le-\sum_i^n x_{ji\alpha}^{(L)},\ \ j\in J_1$$

$$-\sum_i^n x_{ji\alpha}^{(R)} w_i^{(R)}-w_j\le-\sum_i^n x_{ji\alpha}^{(R)},\ \ j\in J_2,J_3$$

$$\sum_i^n x_{ji\alpha}^{(R)} w_i^{(L)}+w_j\le\sum_i^n x_{ji\alpha}^{(R)},\ \ j\in J_1,J_2$$

$$\sum_i^n x_{ji\alpha}^{(L)} w_i^{(L)}+w_j\le\sum_i^n x_{ji\alpha}^{(L)},\ \ j\in J_3$$

$$j\in\{0,\cdots,m\}=J_1\cup J_2\cup J_3,$$
$$J_1\cap J_2=\phi,\quad J_2\cap J_3=\phi,\quad J_3\cap J_1=\phi, \quad (34)$$

ここで、制約式(8)(右辺はy_α^R)に対応する双対変数を$w_i^{(R)}$、制約式(9)(右辺はy_α^L)に対応する双対変数を$w_i^{(L)}$で表すと双対変数は以下のように非負になる[2]。

$$w_i^{(R)}\ge 0,\ w_i^{(L)}\ge 0 \quad (35)$$

感度分析[11,15]により、制約式右辺の許容幅を求める。許容幅には、制約式右辺の値を減少させる方向と増大させる方向があり、本論文では減少方向の値、すなわち制約関係を緩和する方向の制約式右辺の許容幅を用いる。許容幅は、制約式右辺を変化させたときに基底変数の組が変化しない最大の範囲と定義される[11,15]。また、許容幅は、一つの制約式を動かすことが前提であり、他の条件は一定でなければならない[11,15]。

最適解の基底変数ベクトルを$\overline{A}_B$、最適解における基底行列Bの逆行列を$B^{-1}=(\beta_{ij});m\times m$型とすると、$i$番目の制約式の許容幅(許容減少値)は、次の数式により求められる[11]。

$$\Delta y_i=\left|\begin{cases}\max\limits_{k:\beta_{ki}>0}\left\{-\dfrac{a_{Bi}{}^0}{\beta_{ki}}\right\}\\ -\infty,\ if\ \beta_{ki}\le 0(k=1,\cdots,m)\end{cases}\right| \quad (36)$$

ここで、$a_{Bi}{}^0$は、第i番目の基底変数の最適解の値である。この制約関係を緩和させる方向の許容幅は、本手法では周辺データのように等号関係にある制約式に限定して求めている。

式(8)(右辺は$y_{i\alpha}^R$)に対応する許容幅を、

$$\Delta y_i^{(R)}\ge 0 \quad (37)$$

式(9)(右辺は$y_{i\alpha}^L$)に対応する許容幅を、

$$\Delta y_i^{(L)}\ge 0 \quad (38)$$

で表す。ただし、括弧のついた$(R),(L)$に、上限や下限という意味はなく、それぞれ制約式(8)、(9)の右辺のy_α^R, y_α^Lに対応するという意味である。

付録2 Step4 の操作が正当な手続きであることの証明：(ゼロでない双対変数と許容幅(制約式右辺)がある制約式に存在するとき、その制約式に対応するデータ番号の他方の制約式右辺には許容幅以上のスラックが存在することの証明。ただし、許容幅は制約式右辺の制約を緩める方向である)

最適解におけるi^*番目のデータは、定義より、

$$z_{i^*\alpha}^R \geq z_{i^*\alpha}^L,\ y_{i^*\alpha}^R \geq y_{i^*\alpha}^L \tag{39}$$

$$z_{i^*\alpha}^R - z_{i^*\alpha}^L = k_1 \quad (\geq 0) \tag{40}$$

$$y_{i^*\alpha}^R - y_{i^*\alpha}^L = k_2 \quad (\geq 0) \tag{41}$$

ただし、$k_1, k_2 \geq 0$は任意の実数。

したがって、最適解における予測値の下限が等号関係にあるとき、スラック変数を導入して不等号を等号に直すと、

$$-\hat{z}_{i^*\alpha}^L = -y_{i^*\alpha}^R,\ \hat{z}_{i^*\alpha}^R \geq y_{i^*\alpha}^L \tag{42}$$

$$\hat{z}_{i^*\alpha}^L + y_{i^*\alpha}^R = s_{i^*}^{(L)},\ (s_{i^*}^{(L)} = 0) \tag{43}$$

$$\hat{z}_{i^*\alpha}^R - y_{i^*\alpha}^L = s_{i^*}^{(R)},\ (s_{i^*}^{(R)} \geq 0) \tag{44}$$

$\hat{z}_{i^*\alpha}^L$は最適解における予測値の下限、$\hat{z}_{i^*\alpha}^R$は最適解における予測値の上限とする。

ここで、許容幅$\Delta y_{i^*}\ (>0)$より小さな値$\varepsilon > 0$だけ以下のように変化させる。

$$y_{i^*\alpha}^{\prime R} = y_{i^*\alpha}^R + \varepsilon \tag{45}$$

この新たな制約条件のもとで最適解を求めると変化した回帰係数による等号関係が依然として成立している(許容幅[11,15]の定義)から、

$$\hat{z}_{i^*\alpha}^{\prime L} = y_{i^*\alpha}^{\prime R} \tag{46}$$

新たな制約条件のもとで最適解における他方の制約式は、式(44)より

$$\begin{aligned} s_{i^*}^{\prime(R)} &= \hat{z}_{i^*\alpha}^{\prime R} - y_{i^*\alpha}^L \quad (\geq 0) \\ &= \hat{z}_{i^*\alpha}^{\prime R} - (y_{i^*\alpha}^R - k_2) \\ &= \hat{z}_{i^*\alpha}^{\prime R} - (\hat{z}_{i^*\alpha}^{\prime L} - \varepsilon - k_2) \\ &\geq \hat{z}_{i^*\alpha}^{\prime R} - (\hat{z}_{i^*\alpha}^{\prime R} - \varepsilon - k_2) = \varepsilon + k_2 \end{aligned} \tag{47}$$

したがって、変化させた以上のスラックが他方の制約式に必ず存在する。このことから、以下の二点が明らかである。相補性定理より、(1)どちらか一方の(等号関係にある)制約式の制約を緩める方向に許容幅が存在するならば、他方の制約式右辺の双対変数はゼロである。(2)等号関係にある制約式の制約を緩める方向へ許容幅が存在するならば、その許容幅の範囲内で、ファジィ数演算にもとづく変化を他方の制約式右辺に同時に行うことができる。

この操作は、ひとつのデータに対応する二本の制約式を同時に取り扱うことに等しい。

ここで、上のケースでは、予測値の下限が等号関係にある場合の証明であるが、逆のケースすなわち、予測値の上限が等号関係にあるときも同様に証明できる。