ショートノート

頑健な可能性回帰[†]

竹澤 邦夫*

従来の可能性回帰は異常値に弱い。そこで、「平行な2本の直線の間にかんりの数のデータがある、という条件の下で、2本の直線の切片の差を小さくする」という方針に基づく可能性回帰アルゴリズムを開発した。この方法は、データの中に占める異常値の割合によらず対応でき、アルゴリズムが単純なため容易に実施でき、結果の解釈がしやすい点に特徴がある。また、この方法を応用すれば、可能性回帰の文脈に沿った、しかも、異常値に対して頑健な LOWESS (LOcally WEighted Scatter plot Smoothing) を実現することができる。

キーワード：可能性回帰, LOWESS, 非線形回帰, 異常値, ロバスト回帰

1. はじめに

回帰を行うにあたって、異常値の存在は大きな問題になる。データの中のわずかな数の異常値がデータ解析を台無しにしてしまい、データの持つ重要な情報を見失わせてしまうことがある。そのため、これまでに様々な対応策が考えられてきた。

直線回帰において異常値に対してロバストな方法としては、中間値を用いる方法^{1),2)}と、データに重みをつける方法³⁾が有名である。また、平滑化にあたって異常値に対する頑健さを考慮した方法としては、局所的な中間値をとる方法¹⁾と、LOWESS (LOcally WEighted Scatter plot Smoothing)^{4),5),6)}の重みを決定する際に中間値を用いる方法⁷⁾がよく知られている。

これらに共通するのは、データに異常値が含まれているときには有効に機能するが、そうでない場合は最小二乗法を用いた回帰の方がよい結果を与えることである。そのため、実際のデータ解析にあたっては、両方の方法を試して結果を比較する、という手順を必要とする²⁾。こうした煩雑さが、これらの異常値に強いアルゴリズムが実用的

に利用されることが少ないことの一因であろう。

そこで、可能性回帰^{8),9),10),11),12)}(ファジィ線形回帰と呼ばれることもあるが、ここで提案する方法は非線形な最適化を用いるので、可能性回帰の一種と呼ぶこととした)の概念を応用した、異常値に強い方法を開発した。可能性回帰はそのままでは結果が異常値に大きく左右されてしまうことは明らかだが、わずかな修正を施すことによって、異常値に対して頑健にすることが可能になったのである。また、この方法は、データの中に異常値が含まれていてもいなくても、異常値の割合がどれだけであっても、パラメータの値の変更だけで容易に対応できるため、汎用性が高い。さらに、この方法を LOWESS (LOcally WEighted Scatter plot Smoothing) に応用すれば、こうした優れた特徴を生かした、一般性の高い平滑化が実現する。

2. 基本概念

データは、独立変数($\{x_i\}$, $(1 \leq i \leq N)$)も従属変数($\{y_i\}$, $(1 \leq i \leq N)$)も1次元のスカラでクリस्पな値をとるものとする。このデータを用いた回帰を行うにあたって、最小二乗法や中間値などを用いた回帰のように、1本の直線が真の値を推定すると考えるのではなく、可能性回帰の文脈に従って、2本の直線の間範囲が曖昧さを持つ

[†] Robust Possibility Regression
Kunio TAKEZAWA

* Division of Changing Earth and Agro-Environment,
National Institute of Agro-Environmental Sciences

現象の本質を示す、と考える。この2本の直線を求めるにあたって、ここでは以下のような方針をとる。

「平行な2本の直線の間にかんりの数のデータがある、という条件の下で、2本の直線の切片の差を小さくする」

これを実現するためには、次のような E_1 を最小化する方法が考えられる。

$$E_1 = (b_1 - b_2)^2 + \gamma \sum_{i=1}^N (\text{sign}(ax_i + b_1 - y_i) - \text{sign}(ax_i + b_2 - y_i)) \quad (1)$$

ここで、 $ax_i + b_1$ と $ax_i + b_2$ が2本の直線を示し、 γ は正の定数であり、 $\text{sign}(\cdot)$ は以下のような関数である。

$$\text{sign}(p) = \begin{cases} 1 & \text{if } p > 0 \\ 0 & \text{if } p = 0 \\ -1 & \text{if } p < 0 \end{cases} \quad (2)$$

式(1)の右辺の第1項は2本の直線の切片の値が近いことを要求し、第2項は2本の直線の間多くのデータが存在していることを要求している。この形式は、対立するふたつの要求の拮抗をふたつの項の和の最小化で表現し、 γ のような正の定数(tuning parameter)でふたつの要求のつり合いを調整する、という線形回帰でしばしば用いられる手法を手がかりとして開発したものである。これによって、 $y = ax + b$ というモデルにおいて、 b にだけ曖昧さを考慮したモデルが実現する。通常の直線(平面)回帰において従属変数にだけノイズが含まれていると考えることは、 b にだけ曖昧さを仮定することと等価である。従って、ここで提案するモデルは回帰の結果が範囲で示されるという点で可能性回帰の方法に従い、切片にだけノイズ(曖昧さ)を考える、という点で従来の回帰分析の方法に準拠したものとと言える。

しかし、 E_1 は a , b_1 , b_2 の変化に対して不連続な値をとるので、準ニュートン法のような、差分を用いた非線形最適化法を用いるには適さない。そこで、 E_1 を以下のような E_2 で近似する。

$$E_2 = (b_1 - b_2)^2 + \gamma \sum_{i=1}^N (f(ax_i + b_1 - y_i) - f(ax_i + b_2 - y_i)) \quad (3)$$

ここでは、 $f(\cdot)$ として以下のような関数を用いた。

$$f(p) = 2 \cdot \arctan(p) / \pi \quad (4)$$

この E_2 を用いて、簡単なシミュレーション・データを用いた計算を行った。シミュレーション・データは以下のような20個である。

$$x_i = i \quad (1 \leq i \leq 20) \quad (5)$$

$$y_i = i + 30 + e_i \quad (1 \leq i \leq 15) \quad (6)$$

$$y_i = 15 \quad (16 \leq i \leq 20) \quad (7)$$

e_i は、0 と 4 の間の値をとる一様乱数である。始めの15個のデータが正常なデータで、残りの5個のデータ異常値であることを想定している。図1は、 $\gamma = 1$, $\gamma = 10$ として、それぞれについて E_2 を最小化して a , b_1 , b_2 を求めた結果である。非線形の最小化には、EXCEL 4.0 for Windows(日本語版)のアドイン・ソフトであるソルバーに備えられている準ニュートン法を用い、いくつかの初期値を使って最適化を行った。図2は、 $\gamma = 100$, $\gamma = 1000$ とした場合である。図1と図2を見ると、 γ の値が小さいときには異常値の影響がほとんどなく、 γ の値を大きくすると、2本の直線が異常値を含めた領域を形成するようになっていくことが分かる。従って、データに異常値が含まれる割合に応じて γ の値を変えれば様々なデータに対処できることになる。

A: $\gamma = 1$ のときの、 $ax_i + b_1$ (下限) と $ax_i + b_2$ (上限)
B: $\gamma = 10$ のときの、 $ax_i + b_1$ (下限) と $ax_i + b_2$ (上限)

図1 E_2 を最小化して得られた領域

また、5つの異常値を除いた場合も、図3に示すように妥当な領域が得られている。つまり、異常値の存在の有無に関わらず、この単純な方法が利用できる。Rousseeuw²⁾の方法は、データに含まれる異常値の割合が50%未満なら異常値に煩わされない直線回帰が可能であることを特徴としている。しかし、このことは、たとえば異常値がデータの10%だった場合は、40%の有用なデータをあまり役立てていないことを意味している。その点、 E_2 を用いれば、 γ の値を変えることによって2本の直線の間に存在するデータの数进行调整することができるため、データに含まれる異常値の割合に応じた回帰が実現できる。もっとも、得られたデータの中に異常値がとれだけ含まれているかは事前には分からないことが多いので、この方法で得られる領域に含まれるデータの数と γ の値との関係などを参考にして、最適な γ の値を決定しなければならない。その際は、 γ の値を少しずつ

A: $\gamma=100$ のときの、 ax_i+b_1 (下限)と ax_i+b_2 (上限)
B: $\gamma=1000$ のときの、 ax_i+b_1 (下限)と ax_i+b_2 (上限)

図2 E_2 を最小化して得られた領域

A: $\gamma=1$ のときの、 ax_i+b_1 (下限)と ax_i+b_2 (上限)
B: $\gamma=10$ のときの、 ax_i+b_1 (下限)と ax_i+b_2 (上限)

図3 E_2 を最小化して得られた領域

大きくし、 γ の値の割にはたくさんのデータが2本の直線の間に含まれているものを最適な結果とする、というような方針が採用されるものと思われるが、そのことを客観的に判断するための統計量などについては、今後、検討を進める必要がある。

また、この方法は非線形な最適化を必要とするので局所的な極小値に収束してしまう危険が伴うため、いくつかの初期値を用いた最適化を行わなければならない。特に、独立変数が多次元の場合は、より煩わしい問題になると思われる。が、この点は遺伝的アルゴリズムを導入すれば容易に解決するであろう。

3. LOWESS への応用

前節で示した方法を LOWESS に応用すれば、可能性回帰の文脈に沿った、異常値に強い平滑化が可能になる。これは、以下の $E_3(X)$ を最小化することによって実現する。

$$E_3(X) = (b_1 - b_2)^2 + \gamma \sum_{i=1}^N (w(X - x_i) \cdot (f(ax_i + b_1 - y_i) - f(ax_i + b_2 - y_i))) \quad (8)$$

ここで、 X は、従属変数の値を必要とする点における独立変数の値である。従来の LOWESS と同様に、一つの独立変数に対する推定を行うために一度ずつの $E_3(X)$ の最小化を必要とする。 $w(X - x_i)$ は、従来の LOWESS における重みと同じ働きをする関数で、ここでは以下のものを用いる。

$$w(X - x_i) = \exp(-q(X - x_i)^2) \quad (9)$$

ここで、 q は正の定数である。

図4は、シミュレーション・データを用いて、 $\gamma=1000$ 、 $q=0.02$ とし、 $X_i=i$ ($1 \leq i \leq 20$) に対して $E_3(X_i)$ を最小化して得られた、 a を a_i 、 b_1 を b_{1i} 、 b_2 を b_{2i} としたときの $\{a_i X_i + b_{1i}\}$ (下限) と $\{a_i X_i + b_{2i}\}$ (上限) とを示したものである。ここで用いたシミュレーション・データは以下のようなものである。

$$x_i = i \quad (1 \leq i \leq 20) \quad (10)$$

$$y_i = 5 \cdot \sin(0.1 \odot i) + e_i \quad (1 \leq i \leq 20) \quad (11)$$

この e_i は、0 から 2 の間の値をとる一様乱数である。

図 5 は、 $y_5 = y_{10} = y_{15} = 80$ として異常値と見なし、他のデータは図 4 のものと同じものを用いて、同様の計算を行った結果である。異常値に左右されない妥当な領域が得られている。

ここでは、 γ と q の値については、いくつかの値を用いた計算を行い、妥当性の高い結果が得られているものを直感的に選んだが、迅速な計算が必要な場合や独立変数が多次元の場合を考えると、これらの値を自動的に最適化するための統計量が必要であろう。

4. おわりに

従来の回帰分析手法の多くは、真の値に正規分布に従うノイズが加わったものがデータであると仮定していた。しかも、データには異常値が含まれていないという暗黙の前提を置いた上で、回帰という言葉が示す通り、できるだけノイズに左右されない真の値を推定することを目的としている。しかし、実際のデータにおいて、こうした前提が成り立っているかどうかは不明であることが多いため、これらの仮定に基づいた理論やソフトウェアが有効に機能しないことがある。これが、回帰分析に対する漠然とした不信感を生み出す原因の一つであろう。それに対し、可能性回帰は、現象そのものが曖昧さを持つことを認め、唯一のクリスプな真の値を得ようとはせず、データの持つ広がり率直な表現を指向しているため、性質のはっきりしないデータに対応するための有効な手段である。

ここで提案した方法は、そうした可能性回帰の特性に異常値に対する頑健さを加え、データに含まれる異常値の割合に応じた回帰を実現したものである。しかも、この方法を LOWESS に応用することによって生み出される方法は、異常値に強いノンパラメトリック平滑化の方法としては、応用範囲の広いものの一つと言える¹³⁾。

図4 異常値がない場合に、 E_3 を最小化して得られた領域

図5 異常値を3つ含む場合に、 E_3 を最小化して得られた領域

また、異常値を除いたデータが存在する領域を図示する、という方法は探索的データ解析の代表的な手法である箱ヒゲ図(Boxplot)^{11,3)}と類似したものと見なすこともできる。可能性回帰と探索的データ解析は、共に従来の最尤法を中心とした回帰分析の実用的な機能に飽きたらず、独自の方向を探ったものなので、それらが似通った方法論に到達したことには十分な必然性を感じられる。

参 考 文 献

- (1) J. W. Tukey: Exploratory Data Analysis, Addison-Wesley Publishing Company, Reading, Menlo Park, London, Amsterdam, Don Mills, Sydney, 1977.
- (2) P. J. Rousseeuw: Least median of squares regression, Journal of American Statistical Association, Vol.79, pp.871-880, 1984.
- (3) L. C. Hamilton: Regression with Graphics: a Second Course in Applied Statistics, Duxbury Press, Belmont, California, 1992.

- (4) H. Müller: Nonparetric Regression Analysis of Longitudinal data, Springer-Verlag, Berlin, Heidelberg, 1988.
- (5) W. S. Cleveland and S. J. Devlin: Locally weighted regression: an approach to regression analysis by local fitting, Journal of American Statistical Association, Vol.83, No.403, pp. 597-610, 1988.
- (6) 竹澤：ノンパラメトリック回帰からみたファジィ推論, 日本ファジィ学会誌, Vol.5, No.6, pp. 1273-1279, 1993.
- (7) W. S. Cleveland: Robust locally weighted regression and smoothing scatterplots, Journal of American Statistical Association, Vol.74, No.368, pp.829-836, 1979.
- (8) 田中：ファジィモデリングとその応用, pp.134-147, 朝倉書店.
- (9) J. Kacprzyk and M. Fedrizzi(Eds.): Fuzzy Regression Analysis, Omnitech Press, Warsaw and Physica-Verlag, Heidelberg, 1992.
- (10) 田中：可能性回帰分析, 日本ファジィ学会誌, Vol.5, No.6, pp.1260-1272, 1993.
- (11) 石淵：ファジィ回帰分析, 日本ファジィ学会誌, Vol.4, No.1, pp.52-60, 1992.
- (12) 坂和, 矢野：ファジィ線形回帰分析と豚肉の需要関数同定問題への応用, 日本ファジィ学会誌, Vol.6, No.1, pp.49-60, 1994.
- (13) F.T. Wang and D.W. Scott: The L_1 method for robust nonparametric regression, Journal of American Statistical Association, Vol.89, No. 425, pp.65-76, 1994.

(1994年4月25日 受付)
(1994年8月8日 再受付)

[問い合わせ先]

〒305 つくば市観音台3-1-1

農業環境技術研究所 地球環境研究チーム

竹澤 邦夫

TEL・FAX: 0298-38-8279

著者紹介

竹澤 邦夫 (たけざわ くにお)

農業環境技術研究所 地球環境研究チーム

1984年 名古屋大学大学院 工学研究科応用物理学専攻前期課程修了。同年より、農林水産省 農業環境技術研究所に勤務。ノンパラメトリック回帰、ファジィ推論などを用いた、地球環境、気象、農業生産などの解析とモデル化の研究に従事。日本ファジィ学会、情報処理学会、応用統計学会、日本計量生物学会、日本農業気象学会などの会員。