

左打ち切りデータの回帰分析のための期待値最大化法

長谷川 洵† 加藤 毅†

† 群馬大学大学院理工学府

1 はじめに

回帰分析は説明変数から目的変数の値を予測する方法としてあらゆる分野で広く使われている。しかし、しばしば観測値の値が小さすぎて測定限界を超えずに不検出になることがある。不検出データにゼロ、測定限界、測定限界の半分などを埋めて解析されることがしばしばあるが、すると解析結果に致命的な誤差が生じることがある。これに対して、本研究では、目的変数が左打ち切りデータであったときにも統一的な枠組みで回帰分析するための期待値最大化法を開発した。

左打ち切りデータに対する最尤推定法として、Cohenの研究 [1] がよく知られている。Cohen は、データの分布モデルとして、正規分布を仮定し、打ち切られたデータの確率を、測定限界までの累積密度であらわした。そして、今日に至るまで打ち切りデータの解析はCohenの尤度関数を基本にした方法が使われてきた [2]。

本稿では、目的変数が左打ち切りデータであった場合に対する回帰モデルを提案する。また、回帰モデルのパラメータの推定のために、打ち切りデータを潜在変数と解釈して、EM法を与える。

2 打ち切りデータのための最尤推定

2.1 最小二乗推定と最尤推定の関係

まず、通常回帰問題から議論する。訓練用データセットが

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_\ell, y_\ell)$$

からなっていたとする。典型的には二乗誤差の和

$$\sum_{i=1}^{\ell} (y_i - \langle \psi(\mathbf{x}_i), \mathbf{w} \rangle)^2$$

を最小化するように回帰係数 \mathbf{w} を求める。ただし、 ψ は特徴抽出器である。これを統計学の枠組みに入れると次のようになる。目的変数 y は、対になる説明変数 \mathbf{x} から

$$y = \langle \mathbf{w}, \psi(\mathbf{x}) \rangle + \epsilon$$

によって生成されるとする。 ϵ は正規分布雑音で $\epsilon \sim \mathcal{N}(0, \beta^{-1})$ とする。すると、 y の密度は、

$$p(y | \mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y; \langle \mathbf{w}, \psi(\mathbf{x}) \rangle, \beta^{-1}) \tag{1}$$

モデルパラメータ $\boldsymbol{\theta} = (\mathbf{w}, \beta)$ の値を最尤推定によって決定するとする。データセット \mathcal{S}_f に対する尤度関数は

$$p(\mathcal{S}_f | \boldsymbol{\theta}) := p(X) \prod_{i=1}^{\ell} p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$$

となる。ただし、 $p(X)$ は説明変数の集合 $X := \{\mathbf{x}_1, \dots, \mathbf{x}_\ell\}$ が生起する確率密度である。

二乗誤差の和は、定数項を除いて、 $-\frac{2}{\beta} \log(p(\mathcal{S}_f | \boldsymbol{\theta}))$ に等しいことを示すことができる。よって、最小二乗推定解と最尤推定解が等しくなることが分かる。このように統計学の枠組みに帰着させることにより、ベイズ推定やMAP推定など統計学によって培われた方法論に拡張することが可能になる。

2.2 尤度関数

本稿では、目的変数が打ち切りデータになっているときの回帰問題を扱う。訓練用データセットのうち、 n_v 個に関しては、目的変数の値が検出限界 u を上回ったために、 n_v 個のデータペア

$$(\mathbf{x}_1^v, y_1^v), \dots, (\mathbf{x}_{n_v}^v, y_{n_v}^v)$$

の値がすべて得られているが、残りの $n_h (= \ell - n_v)$ 個に関しては、目的変数の値が u を下回ったために、 n_h 個のデータペア

$$(\mathbf{x}_1^h, y_1^h), \dots, (\mathbf{x}_{n_h}^h, y_{n_h}^h)$$

に対して、 y_i^h の値は不明で、

$$y_1^h \leq u, \dots, y_{n_h}^h \leq u$$

という情報のみが得られているとする。 y_i^v も y_i^h も (1) によって生起するとすると、 $y_i^h \leq u$ という事象は

$$p(y_i^h \leq u | \mathbf{x}_i^h, \boldsymbol{\theta}) = \int_{-\infty}^u \mathcal{N}(y_i^h; \langle \mathbf{w}, \psi(\mathbf{x}_i^h) \rangle, \beta^{-1}) dy_i^h$$

すると、尤度関数を

$$L_{c,ml}(\theta) := \log p(X) + \sum_{i=1}^{n_v} \log p(y_i^v | x_i^v, \theta) + \sum_{i=1}^{n_h} \log p(y_i^h \leq u | x_i^v, \theta)$$

と与えることができる。

2.3 EM法の概要

EM法は、尤度関数の値が単調増加するように、EステップとMステップを収束するまで繰り返す方法である。Eステップでは、その時点で得られているモデルパラメータ $\theta_t = (\mu_t, \beta_t)$ を使って潜在変数の事後分布 $q^i(y_i^h)$ を計算し、Q関数に含まれる期待値を計算する。Mステップでは、Q関数をモデルパラメータに関して最大化する。

2.4 Eステップ

$\mu^i := \langle w, \psi(x_i^h) \rangle$ および $\xi^i := (u - \mu^i) \sqrt{\beta}$ とおくと、事後分布は

$$q^i(y_i^h) = \begin{cases} \frac{1}{\Phi(\xi^i)} \mathcal{N}(y_i^h \leq u | x_i^h, \theta) & \text{for } y_i^h \leq u, \\ 0 & \text{for } y_i^h > u \end{cases}$$

と与えられる。この事後分布に基づいて $\mathbb{E}_{q_{t+1}}[y_h]$ および $\mathbb{E}_{q_{t+1}}[\|y_h\|^2]$ を計算する。

2.5 Mステップ

Mステップでは、モデルパラメータ $\theta_{t+1} = (\mu_{t+1}, \beta_{t+1})$ を次のように更新する。

$$w_{t+1} := (X X^T)^{-1} (X_v y_v + X_h \mathbb{E}_{q_{t+1}}[y_h])$$

および

$$\beta_{t+1}^{-1} := \frac{1}{n_v + n_h} (\|X_v^T w_{t+1} - y_v\|^2 + \|X_h^T w_{t+1}\|^2 - 2 \langle X_h^T w_{t+1}, \mathbb{E}_{q_{t+1}}[y_h] \rangle + \mathbb{E}_{q_{t+1}}[\|y_h\|^2]).$$

ただし、 $X := [X_v, X_h]$ とし、

$$X_v := [\psi(x_1^v), \dots, \psi(x_{n_v}^v)], \quad X_h := [\psi(x_1^h), \dots, \psi(x_{n_h}^h)]$$

とおいた。

3 実験

本節では、本研究で開発した左打ち切りデータの回帰分析のための期待値最大化法の推定精度を検証するために、従来打ち切りデータの解析のために用いられてい

た、不検出データを削除する、不検出データにゼロ、測定限界、測定限界の半分を埋める手法との比較実験の結果を報告する。

実験に用いたデータセットは、真の回帰係数 $w = [0.5, 15]$ 、真の分散 $\beta^{-1} = 1.0$ とし、説明変数 $x_1 = 0.0, x_2 = 0.5, \dots, x_{21} = 10.0$ に対する目的変数 y_1, \dots, y_{21} を(1)から生成した。測定限界 u は y_1, \dots, y_{21} の中央値とした。

図1に各手法により推定された直線をプロットした。真の直線と比較すると、他の手法と比べてEM法が一番真の直線に近い直線になっている。

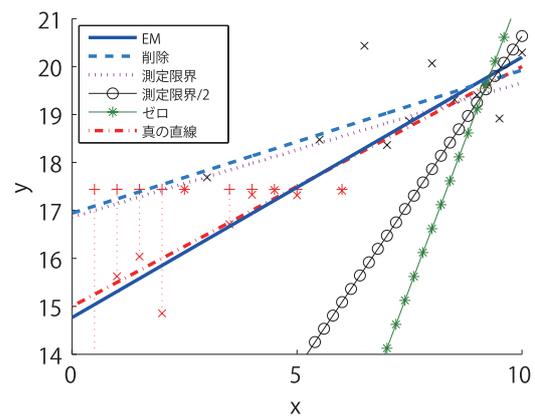


図1 推定された回帰直線

4 おわりに

本研究では、目的変数が左打ち切りデータであったときにも統一的な枠組みで回帰分析するための期待値最大化法を開発した。人工データによる比較実験の結果、従来の手法と比べて精度の向上が確認できた。

謝辞: 本研究は JSPS 科研費 26249075, 40401236 の助成を受けたものである。

参考文献

- [1] Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, pp. 217–237, 1959.
- [2] year. Estimation of concentration ratio of indicator to pathogen-related gene in environmental water based on left-censored data. *Journal of Water and Health*, pp. xx–xx, --. Accepted.