ブートストラップ法とそのロバスト推定への応用

M2011MM051 室梅秀平
指導教員：木村美善

1 はじめに

ロバスト推定量とは, 観測したデータに外れ値が混入していることが予測される場合に有効な推定量である. 通常, 最小二乗法などの一般の推定量においては, 誤差は互いに独立で同一分布に従う対称な確率変数であると仮定して推定を行なう. しかし, その仮定に従わない観測値がデータに含まれていた場合, これらの推定量は良い推定量である根拠を失ってしまう. このように, 仮定したモデルから逸脱する観測値がデータに含まれていた場合でも, その影響を抑え, 良い推定が得られることを目的とした推定量がロバスト推定量である.

ブートストラップ法とは, 観測された限られたデータを元に新たな標本をリサンプリングする手法である. 新しい標本をいくらでもリサンプリングできるので, それらを用いて推定量の偏りや分散, 分布などの統計的性質を大量のデータを用いた反復計算に置き換えて推定することが可能である. 計算機が必要となるが, 複雑な数式を知らなくてもモンテカルロ式に推定を行なえることが大きな魅力である.

ロバスト推定とブートストラップ法はどちらも優れた手法であるが, この二つの手法は相性が良くなく, そのまま組み合わせて使用するといくつかの問題が発生する. これらの問題について考察を行い, その改善案を模索, 検証することが本研究の主な目的である.

また, この問題に対する既知の改善案である Salibian-Barrera and Zamar [2] の Fast Robust Bootstrap(FRB) 法についても考察を行なう.

2 定義

2.1 線形回帰モデル

$(y_i, x_{1i}, \ldots, x_{pi})', i = 1, \ldots, n$ は一般的な分布 H に従う独立な確率ベクトルとする. また, $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{pi})' \in \mathbb{R}^{p+1}$ とする. $\mathbf{x}_i$ によって y_i を説明する次のような線形回帰モデルを考える.

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i, \quad i = 1, ..., n. \tag{1}$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ は回帰パラメータ, $\varepsilon_i \sim N(0, 1)$ は誤差, σ はその尺度である.

外れ値の発生やモデル分布からのずれを考慮に入れるため, 分布 H は興味がある分布 H_0 の ϵ-汚染近傍 $\mathcal{H}_\epsilon$ に属すると仮定する.

$$\mathcal{H}_\epsilon = \{H = (1-\epsilon)H_0 + \epsilon H^* | H^* \in \mathcal{M}\}. \tag{2}$$

ϵ は $0 \leq \epsilon \leq 1/2$ であり, $\mathcal{M}$ は確率分布の全体からなる集合である.

2.2 MM 回帰推定量

ロバスト・ブートストラップ法を適用するロバスト回帰推定量として, MM 回帰推定量 (Yohai [7]) を用いる. MM 回帰推定量は, まず尺度推定値 $\hat{\sigma}_n$ の推定を行い, その推定値を用いて回帰推定値 $\hat{\boldsymbol{\beta}}_n^{MM}$ を推定する. それぞれの推定の式には損失関数 ρ_0, ρ_1 が組み込まれ, この関数が外れ値の重みを低く抑えることでロバストな推定量になっている.

尺度の推定値 $\hat{\sigma}_n$ は次の式 (3) を満たす $\hat{\sigma}_n(\boldsymbol{\beta})$ の内, 最小のものである.

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\hat{\sigma}_n(\boldsymbol{\beta})}\right) = b. \tag{3}$$

$$\hat{\sigma}_n = \min_{\boldsymbol{\beta}} \hat{\sigma}_n(\boldsymbol{\beta}). \tag{4}$$

b は $E_\Phi[\rho_0] = b(\Phi$ は標準正規分布) となるような定数である. なお, このとき $\hat{\sigma}_n = \hat{\sigma}_n(\boldsymbol{\beta})$ となるような $\boldsymbol{\beta}$ を S 回帰推定値と呼び, $\hat{\boldsymbol{\beta}}_n^S$ とする.

次に, $\hat{\sigma}_n$ を用いて回帰推定値 $\hat{\boldsymbol{\beta}}_n^{MM}$ を求める. MM 回帰推定値 $\hat{\boldsymbol{\beta}}_n^{MM}$ は次の式 (5) を満たすものである.

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1'\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_n^{MM}}{\hat{\sigma}_n}\right)\mathbf{x}_i = \mathbf{0}. \tag{5}$$

損失関数 ρ_0, ρ_1 は連続微分可能な偶関数であり, ρ_0 の取り方は漸近効率を, ρ_1 の取り方は破綻点をそれぞれ左右する (Huber [1] を参照).

2.3 ブートストラップ法

ここでは, 経験分布を用いたノンパラメトリックなブートストラップ法について定義する. 観測したデータに基づく経験分布を H_1 とする. この経験分布 H_1 とは, n 個の観測値の各点で確率 $1/n$ を持つ分布である. H_1 から無作為復元抽出された大きさ n の標本をブートストラップ標本と呼び, $(y_i^*, x_{1i}^*, \ldots, x_{pi}^*)', i = 1, ..., n$ と表す. また, $\mathbf{x}_i^* = (1, x_{1i}^*, \ldots, x_{pi}^*)', i = 1, ..., n$ とする.

ブートストラップ標本を元に, 線形モデルの係数 $\boldsymbol{\beta}$ を推定した推定量をブートストラップ推定量と呼び $\hat{\boldsymbol{\beta}}_n^*$ と表す. 前述の MM 回帰推定量の場合, $\hat{\boldsymbol{\beta}}_n^*$ は次の式を満たす値となる.

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1'\left(\frac{y_i^* - \mathbf{x}_i^{*\prime}\hat{\boldsymbol{\beta}}_n^*}{\hat{\sigma}_n^*}\right)\mathbf{x}_i^* = \mathbf{0}. \tag{6}$$

同じく, 尺度のブートストラップ推定量 $\hat{\sigma}_n^*$ は, 次の式を満たす $\hat{\sigma}_n^*(\boldsymbol{\beta})$ の中で最小のものである.

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i^* - \mathbf{x}_i^{*\prime}\boldsymbol{\beta}}{\hat{\sigma}_n^*(\boldsymbol{\beta})}\right) = b. \tag{7}$$

$$\hat{\sigma}_n^* = \min_{\boldsymbol{\beta}} \hat{\sigma}_n^*(\boldsymbol{\beta}). \tag{8}$$

ブートストラップ標本は既知の分布 H_1 から無作為に生成されるため, 何度でも取り直すことが出来る. そのため, ブートストラップ標本を大量に生成してブートストラップ推定量を大量に用意すると, モンテカルロ法の要領で推定量の分布を推定することも出来る (汪他 [6] を参照).

3 ロバスト推定量への適用の際の問題点

このようなブートストラップ法をロバスト推定量に適用する場合, Salibian-Barrera and Zamar [2] によると次のような問題が発生する.

3.1 計算コスト

ブートストラップ法では, 何千何万ものブートストラップ標本を生成し, そのそれぞれに対して推定量の計算を行なう. しかし, 表 1 にあるようにロバスト推定量の計算コストは最小二乗法などと比べて非常に大きく, 大量に反復計算を行なうと計算コストが膨大なものとなってしまう.

表 1 サンプル数 n, 次元数 p の場合の主なロバスト推定量の計算時間 (秒)

(n,p)	$(50,1)$	$(100,1)$	$(50,3)$	$(100,3)$
LS	0.00008	0.00009	0.00008	0.00009
M	0.032	0.055	0.11	0.21
S	0.42	0.92	1.04	2.21
MM	0.45	0.97	1.15	2.43
τ	0.45	0.94	1.31	1.88

3.2 ブートストラップ標本の外れ値の割合

ブートストラップ標本は, 元の標本の経験分布から無作為復元抽出されたものである. ロバスト推定を適用するからには元の標本には外れ値が含まれていることが予想されるが, 元の標本に含まれるどの観測値も平等に抽出され得るため, 外れ値も正常な値もどちらも平等に選ばれ得る. また, 復元抽出であるため 1 つのブートストラップ標本に同じ外れ値が複数抽出されることも考えられる. その結果, ブートストラップ標本に含まれる外れ値の割合は一定にならず, 外れ値が多く含まれ過ぎているブートストラップ標本も生成されてしまう.

ロバスト推定は外れ値を含むデータに対しても良い推定を行なう方法であるが, あまりに外れ値が多く含まれているとロバスト推定であっても良い推定にはならない. 外れ値が多く含まれ過ぎてしまったブートストラップ標本から算出されたブートストラップ推定値は信頼できない値となってしまう. 図 1 は外れ値を含んだサンプルデータのロバスト推定量に対してブートストラップ法を適用した結果であるが, 良いと思われる推定値の付近 (A) から離れたところ (B) にも多くの推定値が流れてしまっている. これらは, 外れ値が含まれ過ぎたブートストラップ標本によるものである.

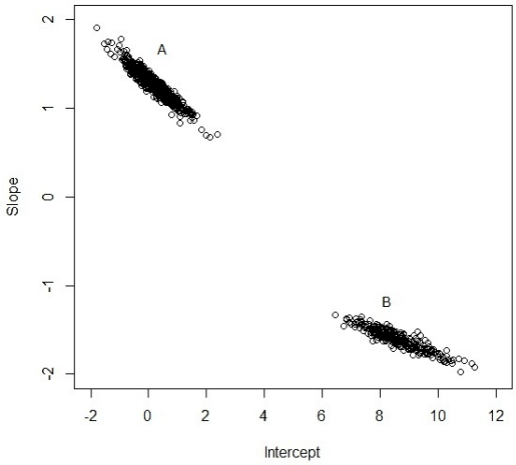

図 1 外れ値が混入したデータへの MM 推定量 $\hat{\boldsymbol{\beta}}_n^{MM}$

4 問題点への改善案

これらの問題点に対する改善案を模索したところ, ロバスト推定量のブートストラップ推定量を算出する際の計算アルゴリズムを一部簡略化することで問題点を改善できることが分かった.

ロバスト推定量を計算する際の一般的な計算アルゴリズムは, Salibian-Barrera, Willems and Zamar [4] によると, 大まかには次のようになる.

1. 解となる $\hat{\boldsymbol{\beta}}_n$ の候補となる $\boldsymbol{\beta}$ を N 個用意する.

2. 用意した候補のそれぞれに対し, その候補の近くにある局所最適解を探し出す.

3. 見つけた局所最適解をそれぞれ比較し, 最も良いものを解となる推定量とする.

手順 1 で複数の候補を用意する理由は, 非凸の最適化問題だからである. 非凸の最適化問題の場合, 最適解に辿り着くためには最適解にある程度近い位置の初期解が必要となるが, それが何処にあるかが分からないため複数用意して虱潰しにする必要がある. 用意すべき候補の数 N はモデルの次数と, 「最適解に辿り着く初期解が N に含まれる確率」をどれだけ必要とするかで次のように変わってくる.

表 2 モデルの次数 p に対して用意すべき係数の候補の数

$p+1$	2	3	5	10
95%	10	22	94	3066
99%	16	34	145	4713

手順 2, 3 は解の候補の数 N の分だけ計算が必要となるため, N の増加と比例して計算コストは増大する.

そこで, 考案したのが, N の数を減らす方法である. N が複数必要なのは最適解が何処にあるかが分からないためであるが, ブートストラップ推定量の最適解 $\hat{\boldsymbol{\beta}}_n^*$ は元の

データに対する推定量の最適解 $\hat{\boldsymbol{\beta}}_n$ の近くにあると考えられる. それならば, ブートストラップ推定量の計算の際には N は $\hat{\boldsymbol{\beta}}_n$ 一つだけで良いのではないかと考えた.

また, このように初期解を一つに絞ることはもう 1 つの問題点に対する対策にもなる. 図 2 は, 外れ値を含むデータに対してロバスト推定を行なう際に, $\boldsymbol{\beta}$ の変動に対して目的関数がどう変動するかを調べたものである.

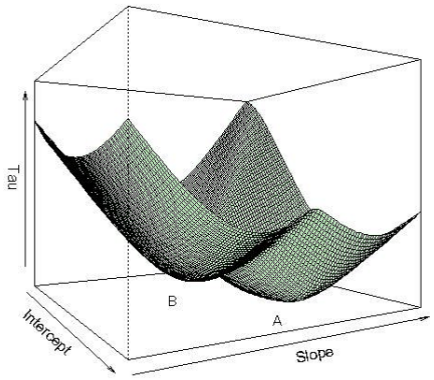

図 2 $\boldsymbol{\beta}$ に対する目的関数の変動

最適解は A であるが, B にも局所最適解が出来ている. 次に, このデータに対するブートストラップ標本で, 外れ値が含まれ過ぎてしまったデータを用意して同じように変動を調べたのが図 3 である.

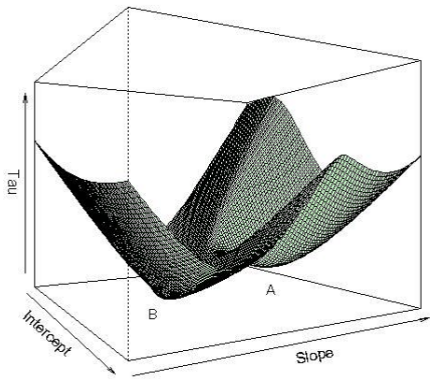

図 3 $\boldsymbol{\beta}$ に対する目的関数の変動

最適解であったはずの A が B より大きくなってしまっている. この時, 通常ならば B が最適解と判断されるが, 提案した方法の場合 A の付近の局所最適解を最適解と判断するため, この場合でも A が推定量として採用される.

図 1 で使ったサンプルデータに対してこの方法でブートストラップを行なうと図 4 のようになった. 図 1 では右下にあった, 外れ値の影響を強く受けたブートストラップ推定量の群がなくなっている. 計算時間も通常のブートストラップ法の場合 517 秒かかっていたのに対し, こちらの手法では 55 秒と 1/10 程度で済んでいる.

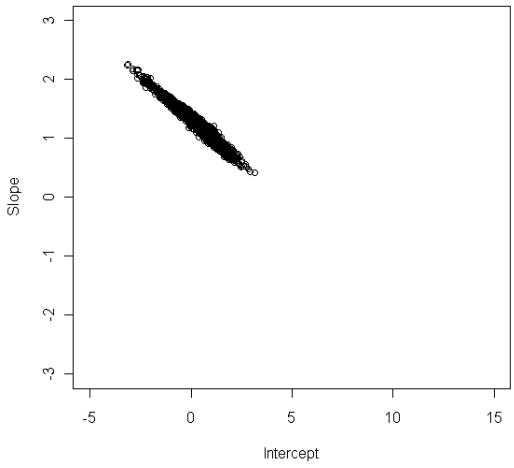

図 4 改良した手法によるブートストラップ推定量

5 Fast Robust Bootstrap

MM 推定量に対してブートストラップ法を適用する際に, 前述の問題点を解決する手法として Salibian-Barrera and Zamar [2] で提案された Fast Robust Bootstrap(FRB) 法がある

この手法では, 先程提案した改善案よりもさらに少ない計算コストでブートストラップを行なうことが可能である. (棚瀬 [5] 参照)

FRB 推定量 $\hat{\boldsymbol{\beta}}_n^{R*}$ では, 先に挙げた問題点を解決するために, ブートストラップ推定量の算出過程を簡略化している. まず, 次のような重み関数を定義する.

$$\omega_i = \frac{\rho_1^{'}(r_i/\hat{\sigma}_n)}{r_i} \tag{9}$$

なお, $r_i = y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}$ である. これを用いて MM 推定量の式 (5) を整理すると次のようになる.

$$\hat{\boldsymbol{\beta}}_n^{MM} = \left[\sum_{i=1}^{n} \omega_i \mathbf{x}_i \mathbf{x}_i^{'}\right]^{-1} \sum_{i=1}^{n} \omega_i \mathbf{x}_i y_i \tag{10}$$

これは ω_i を重み関数とした重み付き最小二乗法の形であり, 非凸の線形最適化問題であった MM 推定量の算出はただの線形問題になる. その結果, 計算コストは大幅に軽減されることになる.

しかし, 実際には ω_i の中に変数 $\boldsymbol{\beta}$ が含まれているため, この問題は線形問題になってはいない. そこで, FRB 法では ω_i の中の $\boldsymbol{\beta}$ を変数ではなく定数に置き換えることでこの問題を線形問題とする.

具体的には, ブートストラップ推定量 $\hat{\boldsymbol{\beta}}^*$ を繰り返し計算する過程において, ω_i の中の $\boldsymbol{\beta}$ を元の観測値から得られた推定量 $\hat{\boldsymbol{\beta}}_n^{MM}$ で代用する.

$$\hat{\boldsymbol{\beta}}_n^{*} = \left[\sum_{i=1}^{n} \omega_i^{*} \mathbf{x}_i^{*} \mathbf{x}_i^{*'}\right]^{-1} \sum_{i=1}^{n} \omega_i^{*} \mathbf{x}_i^{*} y_i^{*} \tag{11}$$

ここで, $r_i^* = y_i^* - \mathbf{x}_i^{*'}\hat{\boldsymbol{\beta}}_n^{MM}$ である.

これにより, 計算コストを大幅に削減することが出来るが, 一部の変数を置き換えたことによる誤差が生じる. それを是正するため, $\mathbf{K}_n$ を定義する.

$$\mathbf{K}_n = \hat{\sigma}_n \left[\sum_{i=1}^{n} \rho_1^{''}(r_i/\hat{\sigma}_n, \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{'} \right]^{-1} \sum_{i=1}^{n} \omega_i \mathbf{x}_i \mathbf{x}_i^{'} \quad (12)$$

この $\mathbf{K}_n$ は元のデータに対してのみ計算し, ブートストラップ推定量を算出する過程で再計算を行う必要はない. これを用いて, 最終的な FRB 推定量 $\hat{\boldsymbol{\beta}}_n^{R*}$ は次のようになる.

$$\hat{\boldsymbol{\beta}}_n^{R*} = \hat{\boldsymbol{\beta}}_n^{MM} + \mathbf{K}_n(\hat{\boldsymbol{\beta}}_n^{*} - \hat{\boldsymbol{\beta}}_n^{MM}) \quad (13)$$

このように計算式を簡略化することで非凸の最適化問題をただの線形問題へと変換し, 計算コストを大幅に削減することが出来る. また, これによりもう 1 つの問題点も同時に改善される. 重み関数 ω_i^* の中で, $\boldsymbol{\beta}$ は元のデータに対して算出されたものをそのまま使っている. そのため, ブートストラップ標本においてリサンプリングされた一つ一つの観測値に対し, 重み関数 ω_i^* が与えるウエイトは, 元のデータにおけるそれと同じになる. すなわち, 元のデータに対する分析において外れ値と判断された観測値はブートストラップ標本においても常に外れ値として扱われ, 推定量の算出に対してほとんど影響を及ぼすことが出来ない. 従って, ブートストラップ標本に外れ値が多く含まれすぎてしまった場合でも, FRB 推定量 $\hat{\boldsymbol{\beta}}_n^{R*}$ は良い推定であり続ける.

図 1, 図 4 と同じサンプルデータへの MM 推定量に対して, FRB 法によるブートストラップを行なった結果が図 5 である. 外れ値の影響を受けすぎた推定量は見られない. 計算時間については 2 秒程度しか要しておらず, 提案した改善案と比較しても段違いに速い.

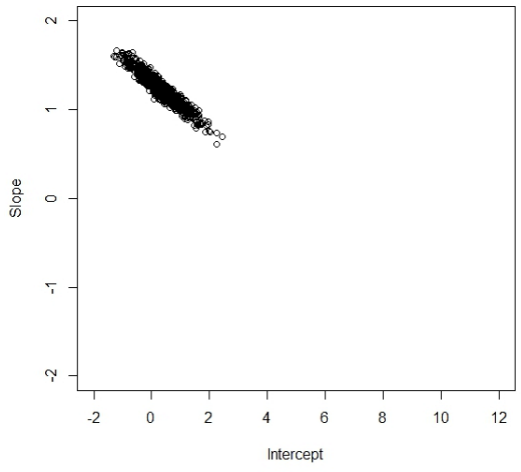

図 5 FRB 法によるブートストラップ推定量

なお, FRB 法の初出は Salibian-Barrera and Zamar [2] であるが, この時に提案された FRB 推定値の導出式と, それより後に発表された論文である Salibian-Barrera and Aelst [3] における FRB 推定値の導出式には違いがあり, 改良がなされていた. ここで紹介した導出式は新しいほうのものである.

6 おわりに

本研究では, ロバスト推定量に対してブートストラップ法を適用する際の問題点について考察し, その改善案についても模索した. 拙いながらも自身による改善案を提案したが, 既に提案されている FRB 法を超えるものとは言い難いものとなった. 提案したブートストラップは FRB 法とは異なるアプローチに基づいたものであったが, この方向よりも FRB 法の方向へと研究を深めていくべきと感じた.

また, 他に改善案を考えるとするならば, ブートストラップ法自体を工夫する手段もあると考えている. 本研究では, ブートストラップ標本をリサンプリングする際には最も一般的な手法を土台にして研究を行なっていたが, ブートストラップ法には様々なバリエーションが存在する. 例えば, 元のデータに対して最初にロバスト推定を行なった段階で, 外れ値と見做される観測値はブートストラップ標本のリサンプリング時には除かれるようにすれば, ブートストラップ標本に含まれる外れ値の割合など気にしないで済むだろう.

研究結果としては少々中途半端な結果となったことが心残りではあるが, また機会があればもう少し研究を進めてみたい.

参考文献

[1] Huber, P.J.(1981). *Robust Statistics,* Wiley, New York.

[2] Salibian-Barrera, M. and Zamar, R.H.(2002). Bootstrapping robust estimates of regression, *The Annals of Statistics*, **30**, 556-582.

[3] Salibian-Barrera, M. and Aelst, S.V.(2008). Robust model selection using fast and robust bootstrap, *Computational Statistics and Data Analysis*, **52**, 5121-5135.

[4] Salibian-Barrera, M., Willems, G. and Zamar, R.(2008). The Fast-τ Estimator for Regression, *Journal of Computational and Graphical Statistics*, **17**, 659-682.

[5] 棚瀬暁俊 (2008). 線形回帰における MM-回帰推定量とロバスト・ブートストラップ法, 南山大学大学院数理情報研究科 修士論文.

[6] 汪金芳・大内俊二・景平・田栗正章 (1992). ブートストラップ法 -最近までの発展と今後の展望-, 行動計量学, **19-2**, 50-81.

[7] Yohai, V.J.(1987). High breakdown-point and high efficiency robust estimates for regression, *The Annals of Statistics*, **15**, 642-656.