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

M2006MM033 棚瀬 暁俊
指導教員　木村　美善

1　はじめに

Salibinan-Barrera and Zamar(2002) は線形回帰において高効率と高破綻点を同時に達成する MM-回帰推定量 ([6] 参照) の漸近分布を推定するための方法として, ロバスト・ブートストラップ法を提案した. この方法は, 従来のブートストラップ法が抱える問題点 (外れ値に対して不安定であること, 計算時間がかかること, ブートストラップ標本ごとに残差の尺度推定値を計算しなければならないこと) を克服する優れたものである. 本論文の目的は,MM-回帰推定量について考察するとともに, シミュレーションを通して, ロバスト・ブートストラップ法の特徴と有効性を明らかにすることである.

2　定義

$(y_1, \mathbf{z}_1')', \ldots, (y_n, \mathbf{z}_n')'$ は互いに独立で同一の分布 H に従う確率変数であり, $\mathbf{x}_i = (1, \mathbf{z}_i')' \in \mathbb{R}^p$ とする. 線形回帰モデル

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_0 + \sigma_0\varepsilon_i, i = 1, ..., n. \tag{1}$$

$y_i \sim F_0, \mathbf{z}_i \sim G_0, (y_n, \mathbf{z}_n')' \sim H_0$ が近似的に成り立ち,y_i と $\mathbf{z}_i$ が独立であると仮定する. ここで F_0 は原点対称で単峰な密度関数 f_0 をもつとする. 外れ値の発生とモデル分布とのずれを考慮に入れるために, データの真の分布 H が H_0 の ε-汚染近傍

$$\mathcal{H}_\varepsilon = \{H = (1-\varepsilon)H_0 + \varepsilon H^*\} \tag{2}$$

に属していると仮定する. ここで $0 \le \varepsilon_0 < 1/2$. H^* は任意の分布とする.

MM-回帰推定量は 2 つの損失関数 ρ_0, ρ_1 に基づいており, それぞれは破綻点と効率に関係するものである. MM-回帰推定量 $\hat{\boldsymbol{\beta}}_n$ は

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1'\Big(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n}\Big)\mathbf{x}_i = \mathbf{0} \tag{3}$$

を満たし,$\hat{\sigma}_n$ は尺度の S-推定量, すなわち $\hat{\sigma}_n$ は等式

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\Big(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\hat{\sigma}_n(\boldsymbol{\beta})}\Big) = b \tag{4}$$

によって定義される M-尺度 $\hat{\sigma}_n(\beta)$ を最小にするものである. ここで $b \in (0,1)$ は定数. また,$\tilde{\boldsymbol{\beta}}_n$ は S-回帰推定量

$$\tilde{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\hat{\sigma}_n(\boldsymbol{\beta}) \tag{5}$$

を表すものとする ([6] 参照).

3　ロバスト · ブートストラップ法

$\hat{\boldsymbol{\beta}}_n$ は (3) を満たす MM-回帰推定量とする. 尺度推定量 $\hat{\sigma}_n$ は (4) から得られる. ここで残差 $r_i = y_i - \hat{\boldsymbol{\beta}}_n'\mathbf{x}_i$,$\tilde{r}_i = y_i - \tilde{\boldsymbol{\beta}}_n'\mathbf{x}_i$ について考える. $\hat{\boldsymbol{\beta}}_n$ と $\hat{\sigma}_n$ は加重最小二乗法によるあてはめの結果としてきちんと表現できる. 重み $\omega_i, \upsilon_i, i = 1, \ldots, n$ は

$$\omega_i = \frac{\rho_1'(r_i/\hat{\sigma}_n)}{r_i} \tag{6}$$

$$\upsilon_i = \frac{\hat{\sigma}_n}{nb}\frac{\rho_0(\tilde{r}_i/\hat{\sigma}_n)}{\tilde{r}_i} \tag{7}$$

とする. このとき次の (3),(4) の加重平均表現が簡単な計算によって得られる.

$$\hat{\boldsymbol{\beta}}_n = \Big[\sum_{i=1}^{n}\omega_i\mathbf{x}_i\mathbf{x}_i'\Big]^{-1}\sum_{i=1}^{n}\omega_i\mathbf{x}_i y_i \tag{8}$$

$$\hat{\sigma}_n = \sum_{i=1}^{n}\upsilon_i(y_i - \tilde{\boldsymbol{\beta}}_n'\mathbf{x}_i) \tag{9}$$

いま,$\{(y_i^*, \mathbf{x}_i^{*\prime})', i = 1, \ldots, n\}$ は観測値からのブートストラップ標本である. $\hat{\boldsymbol{\beta}}_n^*$,$\hat{\sigma}_n^*$ を

$$\hat{\boldsymbol{\beta}}_n^* = \Big[\sum_{i=1}^{n}\omega_i^*\mathbf{x}_i^*\mathbf{x}_i^{*\prime}\Big]^{-1}\sum_{i=1}^{n}\omega_i^*\mathbf{x}_i^* y_i^* \tag{10}$$

$$\hat{\sigma}_n^* = \sum_{i=1}^{n}\upsilon_i^*(y_i^* - \tilde{\boldsymbol{\beta}}_n'\mathbf{x}_i^*) \tag{11}$$

によって定義する. ここで,$\omega_i^* = \rho_1'(r_i^*/\hat{\sigma}_n)/r_i^*$, $\upsilon_i^* = \hat{\sigma}_n\rho_0(\tilde{r}_i^*/\hat{\sigma}_n)/nb\tilde{r}_i^*$, $r_i^* = y_i^* - \hat{\boldsymbol{\beta}}_n'\mathbf{x}_i^*$,$\tilde{r}_i^* = y_i^* - \tilde{\boldsymbol{\beta}}_n'\mathbf{x}_i^*, 1 \le i \le n$. 推定量 $\hat{\boldsymbol{\beta}}_n$,$\hat{\sigma}_n$ および $\tilde{\boldsymbol{\beta}}_n$ は各ブートストラップ標本から再計算されないことに注意する. また,

$$\mathbf{M}_n = \hat{\sigma}_n\Big[\sum_{i=1}^{n}\rho_1''(r_i/\hat{\sigma}_n)\mathbf{x}_i\mathbf{x}_i'\Big]^{-1}\sum_{i=1}^{n}\omega_i\mathbf{x}_i\mathbf{x}_i'$$

$$\mathbf{d}_n = a_n^{-1}\Big[\sum_{i=1}^{n}\rho_1''(r_i/\hat{\sigma}_n)\mathbf{x}_i\mathbf{x}_i'\Big]^{-1}\sum_{i=1}^{n}\rho_1''(r_i/\hat{\sigma}_n)r_i\mathbf{x}_i$$

$$a_n = \hat{\sigma}_n^2\frac{1}{nb}\Big[\sum_{i=1}^{n}\rho_0'(\tilde{r}_i/\hat{\sigma}_n)\tilde{r}_i/\hat{\sigma}_n\Big]$$

とする. このとき, 再計算される $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$ は

$$\hat{\boldsymbol{\beta}}_n^{R*} - \hat{\boldsymbol{\beta}}_n = \mathbf{M}_n(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n) + \mathbf{d}_n(\hat{\sigma}_n^* - \hat{\sigma}_n) \tag{12}$$

によって与えられる ([3],[4] 参照).

3.1 データ解析

タバコの消費量のデータを扱う ([5] 参照). このデータは標本数 $n = 51$ であり, 説明変数を Income(x_1),Price(x_2),Age(x_3), HS(x_4) として, 応答変数を Sales(y) としよう. ただし,

Sales　州において販売されたタバコのパック数
Income　州における一人当たりの収入 (ドル単位)
Price　州におけるタバコ 1 パックの重みつき平均価格 (セント単位)
Age　州に住む人の年齢の中央値
HS　州において最終学歴が高卒で 25 才以上の割合

このとき線形回帰モデル (1) における MM-回帰推定値および LS-推定値を求めると

$$\hat{\beta}_n^{MM} = (102.735, 0.020, -1.570, 0.324, -0.297)$$
$$\hat{\beta}_n^{LS} = (88.439, 0.022, -3.408, 3.669, -0.460)$$

が得られる. ここでは, ロバスト・ブートストラップ法および従来のブートストラップ法により線形回帰モデルの仮説 $H_0 : \beta = 0$ の検定を行う. リサンプリング回数は 10000 回とし, 信頼区間は対称に両側 95%,99% とする. 各推定量のブートストラップ信頼区間は次の表 1 で与えられる.

robust bootstrap				
β	95%CI		99%CI	
β_0	41.942	189.699	24.968	229.720
β_1	0.008	0.030	0.005	0.034
β_2	−4.244	−1.037	−4.883	−0.472
β_3	−0.322	6.079	−1.198	7.240
β_4	−1.455	0.248	−1.746	0.501
bootstrap for MM-estimates				
β	95%CI		99%CI	
β_0	25.044	342.244	−0.272	386.125
β_1	0.005	0.042	−0.001	0.046
β_2	−4.982	−0.750	−5.439	0.018
β_3	−3.580	7.141	−5.020	8.687
β_4	−2.199	0.334	−2.456	0.833
bootstrap for LS-estimates				
β	95%CI		99%CI	
β_0	−27.425	193.539	−72.951	224.713
β_1	0.007	0.037	0.002	0.041
β_2	−5.727	−1.189	−6.521	−0.037
β_3	0.002	8.155	−1.195	9.956
β_4	−1.437	0.766	−1.715	1.235

表 1　原標本におけるブートストラップ信頼区間

ロバスト・ブートストラップにおいては両側 5%,1% のとき $\beta_3 = 0, \beta_4 = 0$ の仮説は棄却されないが $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0$ においては棄却されることになる. MM-推定量に対するブートストラップ (以下,BMM) は 95% 信頼区間に対しては同様の結果が得られるが,99% 信頼区間に対しては $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0$ は棄却されず 95% 信頼区間のときと異なる結果になる. 推定結果では $\hat{\beta}_2^{MM} = -1.570$ であるため BMM の 99% 信頼区間では適切な検定ができないのかもしれない. これは推定量のブートストラップ分布の正規近似の良さを見れば明らかである. ここでは例として $\hat{\beta}_2^{*R}$(図 1), $\hat{\beta}_2^{*MM}$(図 2), $\hat{\beta}_2^{*LS}$(図 3) を挙げておく. 一方 LS-推定量に対するブートストラップ (以下,BLS) においては 95% で $\beta_3 = 0$ が棄却されるのに対し,99% では棄却されない. また, 切片 β_0 では 95%,99% でともに棄却されない.

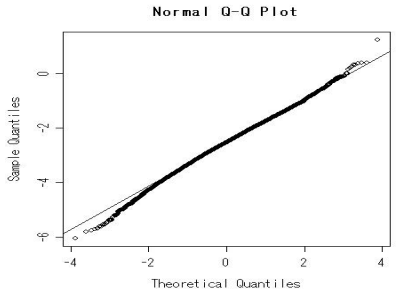

図 1　robust bootstrap qq-plot

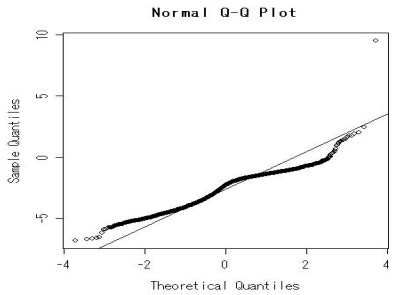

図 2　bootstrap for MM-estimates qq-plot

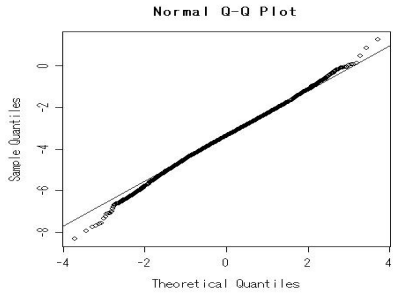

図 3　bootstrap for LS-estimates qq-plot

次に MM-回帰推定量による標準化残差から外れ値と思われる観測値を削除し, 再度, 解析を行うと, 各推定量 $\hat{\beta}_n^{MM}, \hat{\beta}_n^{LS}$ は次で与えられる.

$$\hat{\beta}_n^{MM} = (92.569, 0.017, -1.498, 1.001, -0.224)$$
$$\hat{\beta}_n^{LS} = (91.805, 0.015, -1.466, 1.126, -0.188)$$

$\hat{\beta}_n^{LS}$ は外れ値に引っ張られてしまっているため $\beta_3 = 3.669$ であり, 外れ値除去後のデータの推定値とは大き

く値が異なる. ゆえに, 前述のとおり 95% 信頼区間において $\beta_3 = 0$ の仮説が棄却されないという結果になるのであろう.

次の表 2 は, 外れ値除去後のブートストラップ信頼区間である.

robust bootstrap				
β	95%CI		99%CI	
β_0	42.399	157.417	29.193	184.412
β_1	0.009	0.024	0.007	0.027
β_2	−2.303	−0.740	−2.605	−0.421
β_3	−0.978	3.155	−1.584	3.833
β_4	−0.806	0.308	−1.028	0.490
bootstrap for MM-estimates				
β	95%CI		99%CI	
β_0	29.132	314.336	−4.186	335.209
β_1	0.008	0.035	−0.001	0.038
β_2	−3.496	−0.514	−4.300	1.196
β_3	−3.752	4.417	−5.401	6.182
β_4	−1.854	0.427	−2.042	0.919
bootstrap for LS-estimates				
β	95%CI		99%CI	
β_0	49.018	148.146	39.076	171.115
β_1	0.009	0.023	0.007	0.025
β_2	−2.264	−0.717	−2.545	−0.386
β_3	−0.553	2.997	−1.055	3.794
β_4	−0.693	0.291	−0.871	0.431

表 2　外れ値除去後のブートストラップ信頼区間

ロバスト · ブートストラップに関しては原標本, 外れ値除去後の標本のいずれの場合も仮説 $H_0 : \beta = 0$ の検定結果は変わらないだろう. BMM に関しても同様に, 信頼区間の長さは原標本のブートストラップ信頼区間と比べ短くなっているが, 検定結果は変わらない. BLS ではロバスト · ブートストラップと同様の結果となる. しかし, これは原標本のときとは違う結果であり, 外れ値の影響を受けていたことがよくわかる.

4　ロバスト · ブートストラップ検定

ここで扱う検定は前述のロバスト · ブートストラップ法を使用したものである. 以下では, この検定をロバスト · ブートストラップ検定とよび, 位置 · 尺度調整法を用いた手法へと拡張する (Efron,B. and Tibshirani,R.(1993), 汪 金芳, 田栗 正章 (1996), 桜井 裕仁, 高橋 邦彦 (2002) 参照).

定義 1 観測値に対し, 次のような位置 · 尺度変換

$$x_i^\dagger = (x_i - \mu_x^{MM})/\hat{\sigma}_x \quad i = 1, \ldots, m$$
$$y_j^\dagger = (y_j - \mu_y^{MM})/\hat{\sigma}_y \quad j = 1, \ldots, n$$

を行い,$\mathbf{x}^\dagger = \{x_1^\dagger, \ldots, x_m^\dagger\}$, $\mathbf{y}^\dagger = \{y_1^\dagger, \ldots, y_n^\dagger\}$ とする. ただし,μ_x^{MM}, μ_y^{MM} はそれぞれの MM-位置推定量であり,$\hat{\sigma}_x$,$\hat{\sigma}_y$ は S-尺度推定量である.

擬似データ $\mathbf{x}^\dagger, \mathbf{y}^\dagger$ から作られる経験分布を, それぞれ

$$F_m(x) = m^{-1}\sum_{i=1}^{m}\delta(x_i^\dagger \le x), G_n(y) = n^{-1}\sum_{i=1}^{n}\delta(y_i^\dagger \le y)$$

とする. ただし $\delta(\cdot)$ は定義関数を表している.$F_m(x)$ からの無作為標本を $\mathbf{x}^* = \{x_1^{*b}, \ldots, x_m^{*b}\}$,$G_n(y)$ からの無作為標本を $\mathbf{y}^* = \{y_1^{*b}, \ldots, y_n^{*b}\}$ とし, これらに基づいてロバスト · ブートストラップ検定統計量の値

$$t^{*b} = T(\mathbf{x}^*, \mathbf{y}^*) = \frac{\mu_x^{*R} - \mu_y^{*R}}{\sqrt{(\sigma_x^{*R})^2/(m-1) + (\sigma_y^{*R})^2/(n-1)}}$$

を計算する. ただし $\mu_x^{*R}, \mu_y^{*R}, \sigma_x^{*R}, \sigma_y^{*R}$ はそれぞれロバスト · ブートストラップ位置推定量と尺度推定量である.

前述の計算を B 回繰り返し, ロバスト · ブートストラップ p 値のモンテカルロ近似値を次により計算する.

$$\hat{p} = \frac{1}{B}\sum_{b=1}^{B}\delta(t^{*b} \ge t_{obs})$$

ここで $t_{obs} = T(\mathbf{x}, \mathbf{y})$ はもとの観測値に基づく検定統計量の実現値であり

$$T(\mathbf{x}, \mathbf{y}) = \frac{\hat{\mu}_x^{MM} - \hat{\mu}_y^{MM}}{\sqrt{(\hat{\sigma}_x)^2/(m-1) + (\hat{\sigma}_y)^2/(n-1)}} \quad (13)$$

で定義される. 有意水準 α のときロバスト · ブートストラップ検定を次により行う.

$$\begin{cases} \hat{p} > \alpha \rightarrow \text{帰無仮説を採択} \\ \hat{p} \le \alpha \rightarrow \text{帰無仮説を棄却} \end{cases}$$

4.1　データ解析

ここでアルバカーキの 5 つの洋服工場でのごみのデータに関して, ブートストラップ検定を行う. 使用されるデータは 5 つのデータの中で工場 3, 工場 1 のデータを抜粋したものである ([1] 参照). 工場 3(PT3), 工場 1(PT1) をそれぞれ変数 X, Y としよう. このとき,MM-位置推定量 $\hat{\mu}_x^{MM} = 4.919, \hat{\mu}_y^{MM} = 1.159$ である. しかし, 標本平均は $\bar{X} = 4.832, \bar{Y} = 4.523$ である. 変数 X に関して明らかに $\hat{\mu}_x^{MM}$ と $\bar{X}$ の値は異なっているため, データの外れ値が推定値に影響を与えているかもしれない. この 2 つの工場のデータの平均値に差があるか $H_0 : \mu_x - \mu_y > 0$ の検定を行う.

まず, 原標本を用いてロバスト推定量による検定統計量および従来の検定統計量を求める. そして, ブートストラップ検定, ロバスト · ブートストラップ検定を行い, 位置 · 尺度調整法により両側 90%, 95%, 99% 信頼区間を求める.

ブートストラップ分布の裾は外れ値により影響を受けており (図 5), 正確な信頼区間を構成していないと思われる. この場合は同時に検定統計量も影響を受けていると予想され, これは検定に深刻な問題となる. しかし, ロバスト · ブートストラップ分布の方はより正確な近似分

検定統計量	
robust bootstrap	bootstrap
2.831	0.131

表 3　原標本における検定統計量

robust bootstrap test					
90%CI		95%CI		99%CI	
−2.882	2.041	−3.359	2.507	−4.713	3.553
bootstrap test					
90%CI		95%CI		99%CI	
−1.360	2.709	−1.580	3.341	−2.014	4.524

表 4　原標本におけるブートストラップ信頼区間

布を構成している (図 4).

ロバスト · ブートストラップの検定統計量と信頼区間を照らし合わせると, 両側 95,90% で帰無仮説が棄却され,2 つのデータの平均に差があるといえる. しかしながら, 従来のブートストラップ検定では全ての信頼区間のケースにおいて帰無仮説は棄却されず,2 つの平均は等しいと判断される.

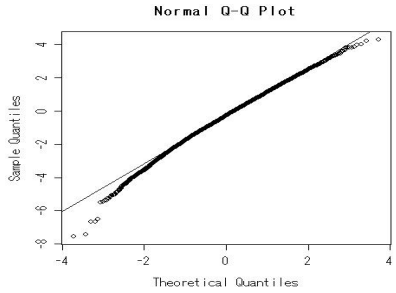

図 4　robust bootstrap test qq-plot

残差から外れ値と思われる観測値を原標本から取り除き再度解析を行う. 表 5,6 はその解析結果である.

いずれの手法においても帰無仮説が棄却され,2 標本の差があるといえるだろう. ロバスト · ブートストラップ検定では外れ値が存在していた原標本による検定結果と変わらないが, 従来のブートストラップ検定では明らかに異なるといえる.

5　おわりに

本論文においては MM-推定量によるロバスト · ブートストラップ法に焦点を当てて研究を行った. これは MM-推定量の特徴である Tukey の ρ 関数を用いていることや各残差に対して重みを付けていることにより, 優れた頑健性をもっている. さらに, 推定量のバイアスに対して補正要素が働くことにより, 頑健性だけでなく正規近似の正確さも表現している. その結果, ブートストラップ法の問題点を克服しており, また変数が高次元の場合の回帰推定の計算量のコストも十分に抑えている.

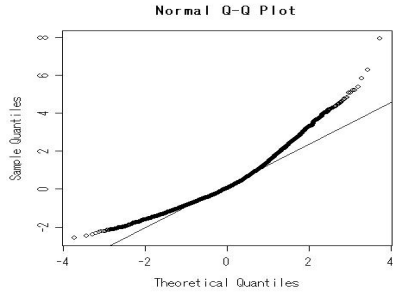

図 5　bootstrap test qq-plot

検定統計量	
robust bootstrap	bootstrap
3.436	3.559

表 5　外れ値除去後の検定統計量

robust bootstrap test					
90%CI		95%CI		99%CI	
−1.511	1.478	−1.773	1.748	−2.305	2.372
bootstrap test					
90%CI		95%CI		99%CI	
−1.601	1.767	−1.916	2.163	−2.524	2.847

表 6　外れ値除去後のブートストラップ信頼区間

リサンプリングされた標本が原標本の外れ値の割合を上回る場合や, その推定量の破綻点を超えてしまうことは, ブートストラップ法の問題点といえる. ロバスト · ブートストラップ法の有効性や 2 標本検定への応用の汎用性を示すことができたのも本研究の成果である.

参考文献

[1] Koopmans,L.(1987). *Introduction to Contemporary Statistical Methods*, Duxbury Press.

[2] Salibian-Barrera,M.(2006a). The asymptotics of MM-estimators for linear regression with fixed designs. *Metrika* **63**,283-294.

[3] Salibian-Barrera,M.(2006b). Bootstrapping MM-estimators for linear regression with fixed designs. *Statictics and Probability Letters* **76**,1287-1297.

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

[5] Chatterjee,S. and Hadi,A.S.(2006). *Regression Analysis by Example*,Wiley,New York.

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