変数選択後の統計的推測

末石，直也

(Citation)
国民経済雑誌,227(2):15-27

(Issue Date)
2023-02-10

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100479007

(URL)
https://hdl.handle.net/20.500.14094/0100479007

ISSN 0387-3129

国民経済雑誌

変数選択後の統計的推測

末　石　直　也

国民経済雑誌　第227巻　第２号　抜刷
2023年２月

神戸大学経済経営学会

『国民経済雑誌』第227巻第2号（2023年2月），15-27

変数選択後の統計的推測

末　石　直　也[a]

本稿では，統計的推測に用いる線形回帰モデルが標本を用いて選択された場合の，統計的推測の問題について考察する。線形回帰モデルに含まれる説明変数が標本に依存して決定されるとき，分析に用いられるモデルがランダムに決定されるため，変数選択後に行われる統計的推測の結果を歪めてしまう。本稿では，このような問題に対処するために近年提案されている信頼区間の構築方法を紹介する。

キーワード　　一様性，信頼区間，変数選択

1　はじめに

本稿では，線形回帰モデルの係数パラメータの信頼区間の構築方法について，従来から用いられている手法の問題点を整理するとともに，それらの問題点を克服するための最近の研究成果を概観する。

以下では，次のような線形回帰モデルを考える。

$$y_i = x_i'\beta + e_i, \quad i=1, \ldots, n$$

ただし，$x_i=(x_{i1}, \ldots, x_{ip})'$ は p 次元の説明変数のベクトル，$\beta=(\beta_1, \ldots, \beta_p)'$ は p 次元の係数のベクトルである。ベクトル x_i はデータの分析者が候補として持っている全ての説明変数を含んでいる。説明変数は確率変数ではなく，定数であるとする。

データの分析者が入手可能なすべての変数を用いて回帰分析を行うことは稀であろう。通常は何らかの方法で必要な変数のみを選択した後に，興味のあるパラメータに関する統計的推測がなされる。変数選択は分析対象のドメイン知識に基づいて行われることもあれば，観察された標本に基づいて行われることもある。また，観察された標本に基づく場合でも，統計学的に根拠のある手続きを用いて変数選択を行うこともあれば，自分の興味のあるパラメータを有意にするために恣意的に変数を選ぶこともあるかもしれない。

変数選択が標本に依存して行われるとき，標本が異なれば，選ばれる変数も異なる可能性がある。標本はランダムに抽出されるので，このことは最終的な分析に用いられるモデルが

a　神戸大学大学院経済学研究科，sueishi@econ.kobe-u.ac.jp

ランダムに決定されるということを意味する。ところが，多くの実証研究においては変数選択の不確実性は無視され，得られたモデルをあたかもあらかじめ与えられた正しいモデルであるかのように扱って，統計的推測が行われている。本稿ではこのような方法の問題点を指摘するとともに，近年考察されている新しい統計的推測の手法を紹介する。

2　従来の方法とその問題点

2.1　変数選択が統計的推測に与える影響

変数選択後の統計的推測の問題を整理するため，Leeb and Pötscher（2005）の例を借用する。以下の例では説明変数の候補が2つしかないが，本質を理解するには十分である。

ある（β_1, β_2）が存在して，y_i は次のようなデータ生成過程から得られるものとする。

$$y_i = x_i'\beta + e_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \qquad (1)$$

ただし，$e_i \sim N(0, \sigma^2)$ であり，σ^2 は既知であるものとする。興味があるパラメータは β_1 のみであり，β_1 の信頼区間を構築するために，x_{i2} を推定する線形回帰モデルに入れるかどうかを決定するという状況を考える。

さらにノーテーションを導入する。x_{i1} のみを説明変数とするモデルを $\mathbb{M}_1$，x_{i1} と x_{i2} の両方を説明変数とするモデルを $\mathbb{M}_2$ で表す。また，正しく定式化されたモデルのうち，最も未知パラメータが少ないものを $\mathbb{M}_0$ という記号で表す。すると，(1)の β_2 の真の値に応じて

$$\mathbb{M}_0 = \begin{cases} \mathbb{M}_1 & \beta_2 = 0 \text{ のとき} \\ \mathbb{M}_2 & \beta_2 \neq 0 \text{ のとき} \end{cases}$$

が成り立つ[1)]。また，次のように行列を定義する。

$$\sigma^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

さらに，$\rho \neq 0$ としておく。

それぞれのモデルを用いた場合の $\beta_j (j=1, 2)$ の OLS 推定量を，$\hat{\beta}_j(\mathbb{M}_1)$ と $\hat{\beta}_j(\mathbb{M}_2)$ で表す。モデル $\mathbb{M}_1$ においては β_2 は推定しないが，$\hat{\beta}_2(\mathbb{M}_1) = 0$ であると考える。β_1 の推定量について，$\mathbb{M}_0 = \mathbb{M}_1$ であれば，

$$\hat{\beta}_1(\mathbb{M}_1) \sim N\left(\beta_1, \frac{\sigma_1^2(1-\rho^2)}{n}\right)$$

が成り立つ。Z 統計量を $Z = \sqrt{n}\,(\hat{\beta}_1(\mathbb{M}) - \beta_1)/(\sigma_1\sqrt{1-\rho^2})$，$z_\alpha$ を標準正規分布の上側 α 点とすると，$P(|Z| \leq z_{\alpha/2}) = 1-\alpha$ であることから，β_1 の信頼水準 $1-\alpha$ の信頼区間は

$$\mathrm{CI}_1(\mathbb{M}_1) = \left[\hat{\beta}_1(\mathbb{M}_1) - z_{\alpha/2}\frac{\sigma_1\sqrt{1-\rho^2}}{\sqrt{n}}, \hat{\beta}_1(\mathbb{M}_1) + z_{\alpha/2}\frac{\sigma_1\sqrt{1-\rho^2}}{\sqrt{n}}\right] \qquad (2)$$

によって得られる。Z 統計量の分布は未知パラメータ β_1 には依存しないので，信頼区間が β_1

を含む確率（被覆確率）もβ_1の値に依存しない。一方，$\mathbb{M}_0=\mathbb{M}_2$のときには

$$\hat{\beta}_1(\mathbb{M}_2)\sim N\left(\beta_1, \frac{\sigma_1^2}{n}\right)$$

が成り立ち，β_1の信頼水準$1-\alpha$の信頼区間は

$$\mathrm{CI}_1(\mathbb{M}_2)=\left[\hat{\beta}_1(\mathbb{M}_2)-z_{\alpha/2}\frac{\sigma_1}{\sqrt{n}}, \hat{\beta}_1(\mathbb{M}_2)+z_{\alpha/2}\frac{\sigma_1}{\sqrt{n}}\right] \qquad (3)$$

によって得られる。この信頼区間の被覆確率も，(β_1, β_2)の値には依存していない。$\mathbb{M}_0=\mathbb{M}_1$のとき，(2)のほうが(3)よりも信頼区間は短くなる。一方，$\mathbb{M}_0=\mathbb{M}_2$であるときに誤って$\hat{\beta}_1(\mathbb{M}_1)$で$\beta_1$を推定してしまうと，欠落変数バイアスが生じてしまい，(2)は適切な信頼区間とはならない。よって，$\mathbb{M}_0$に応じて

$$\mathrm{CI}_1(\mathbb{M}_0)=\begin{cases}\mathrm{CI}_1(\mathbb{M}_1) & \mathbb{M}_0=\mathbb{M}_1 \text{ のとき}\\ \mathrm{CI}_1(\mathbb{M}_2) & \mathbb{M}_0=\mathbb{M}_2 \text{ のとき}\end{cases}$$

のように信頼区間を使い分けることが可能であれば理想的である。

実際にはどちらのモデルが$\mathbb{M}_0$であるかはわからないため，例えば，次のようなルールに基づいて，信頼区間の構築のために用いるモデル$\hat{\mathbb{M}}$を決定することにする。

$$\hat{\mathbb{M}}=\begin{cases}\mathbb{M}_1 & |\sqrt{n}\hat{\beta}_2(\mathbb{M}_2)/\sigma_2|\leq c \text{ のとき}\\ \mathbb{M}_2 & |\sqrt{n}\hat{\beta}_2(\mathbb{M}_2)/\sigma_2|>c \text{ のとき}\end{cases}$$

ただし，$c>0$はデータ分析者が自ら決定する閾値である。$c=z_{\alpha/2}$とすれば，Z検定に対応する。有意水準αで$H_0:\beta_2=0$を検定し，H_0が棄却されれば$\mathbb{M}_2$を用い，そうでなければ$\mathbb{M}_1$を用いる。また，詳しくは論じないが，$c=\sqrt{2}$とすると赤池情報量規準（AIC）に基づく変数選択に対応し，$c=\sqrt{\log n}$とするとベイズ情報量規準（BIC）に基づく変数選択に対応する。

上記のように変数選択を行い，選ばれたモデルに基づいて信頼区間を求めるならば，β_1の信頼区間は次のように表される。

$$\mathrm{CI}_1(\hat{\mathbb{M}})=\begin{cases}\mathrm{CI}_1(\mathbb{M}_1) & \hat{\mathbb{M}}=\mathbb{M}_1 \text{ のとき}\\ \mathrm{CI}_1(\mathbb{M}_2) & \hat{\mathbb{M}}=\mathbb{M}_2 \text{ のとき}\end{cases} \qquad (4)$$

この信頼区間の被覆確率は

$$\begin{aligned}P(\beta_1\in\mathrm{CI}_1(\hat{\mathbb{M}}))=&P(\beta_1\in\mathrm{CI}_1(\mathbb{M}_1)|\hat{\mathbb{M}}=\mathbb{M}_1)P(\hat{\mathbb{M}}=\mathbb{M}_1)\\ &+P(\beta_1\in\mathrm{CI}_1(\mathbb{M}_2)|\hat{\mathbb{M}}=\mathbb{M}_2)P(\hat{\mathbb{M}}=\mathbb{M}_2)\end{aligned}$$

である。ただし，$P(\beta_1\in\mathrm{CI}_1(\mathbb{M}_j)|\hat{\mathbb{M}}=\mathbb{M}_j)$は，モデル$\mathbb{M}_j$が選ばれたという条件の下で，信頼区間が$\beta_1$を含む条件付確率を表す。変数選択の結果は標本に依存し，常に適切なモデルが選ばれるとは限らない。そのため，(4)の被覆確率は一般には$1-\alpha$とは異なり，通常は$1-\alpha$よりも小さくなる。つまり，変数選択によって選ばれたモデルをあたかも正しいモデ

ルであるかのように扱って信頼区間を求めてしまうと，信頼区間の被覆確率は意図した信頼水準よりもずっと小さくなってしまう可能性があるのである。

2.2　一致性を満たす変数選択法

ここでよくある誤解は，一致性を満たす方法で変数選択を行えば，サンプルサイズが十分大きいときには2.1節の最後で議論した問題は生じないというものである。変数選択法が一致性を満たすとは，$n\to\infty$ のとき

$$P(\hat{\mathbb{M}}=\mathbb{M}_0)\to 1$$

が成り立つことをいう。つまり，漸近的には $\mathbb{M}_0$ を正しく選ぶことができることを意味する。現在の例においては，AICは一致性を満たさないが，BICは一致性を満たす。

一致性が満たされるとき，(4)で与えられる信頼区間は，任意の (β_1, β_2) について

$$\lim_{n\to\infty} P(\beta_1\in \mathrm{CI}_1(\hat{\mathbb{M}}))=1-\alpha \qquad (5)$$

を満たすことが示される。すなわち，どちらのモデルが $\mathbb{M}_0$ であろうと，$\mathrm{CI}_1(\hat{\mathbb{M}})$ は漸近的には信頼水準 $1-\alpha$ の信頼区間になっている。

上記の議論に何も間違いはないのだが，(5)が成り立つことは，必ずしも $\mathrm{CI}_1(\hat{\mathbb{M}})$ が有限標本でも良い信頼区間であることを意味しない。このような漸近的な議論は，信頼区間が持つ欠陥を覆い隠してしまう。$\mathrm{CI}_1(\mathbb{M}_2)$ と $\mathrm{CI}_1(\hat{\mathbb{M}})$ の大きな違いは，後者の被覆確率は有限標本において (β_1, β_2) の値に依存する点にある。そのため，$\mathrm{CI}_1(\hat{\mathbb{M}})$ の被覆確率が $1-\alpha$ に十分近くなるために必要なサンプルサイズは，(β_1, β_2) の値に依存して異なる。あるパラメータの値についてはサンプルサイズ100でも $1-\alpha$ に十分近い被覆確率が得られる一方で，別のパラメータの値についてはサンプルサイズ10000でも被覆確率は $1-\alpha$ よりずっと小さいということが起こりうるのである。

このような問題が起こるのは，(5)がパラメータの値についての各点収束の結果であり，一様収束ではないためである。その点を明確にするため，確率を P_{n,β_1,β_2} という記号を用いて表すことにする。これは，被覆確率がサンプルサイズ n とパラメータ (β_1, β_2) の真値に依存するという意味である。変数選択にBICを用いた場合には，(4)について

$$\lim_{n\to\infty}\min_{\beta_1,\beta_2} P_{n,\beta_1,\beta_2}(\beta_1\in \mathrm{CI}_1(\hat{\mathbb{M}}))=0$$

となることが知られている。つまり，サンプルサイズが無限に大きくても，パラメータの値によっては，被覆確率が0になるようなケースが存在するのである。ここでのポイントは，極限を取る前に最小値が取られているところにある。各サンプルサイズについて，そのサンプルサイズの下で最も被覆確率を小さくするようなパラメータを選んでいるので，最小値を達成するパラメータはサンプルサイズに依存して変化する。

一般に，BICに限らず多くの変数選択の方法について，パラメータの値について一様に信

頼水準に近い被覆確率をもたらす信頼区間を構築することは不可能であることが知られている（Leeb and Pötscher 2008a）。したがって，説明変数の候補がそれほど多くない場合には，変数選択などせず，すべての変数をそのまま用いるのが無難である。最初にすべての変数を含むモデルを推定したのち，有意にならなかった変数を省いて推定しなおすようなことは推奨されない。モデルに含まれる説明変数が多いと，信頼区間が長くなってしまうというデメリットがあるものの，正しい被覆確率が得られないことに比べれば，その弊害はずっと小さい。

ちなみに，変数選択後に信頼区間を求める場合には，あるモデル $\mathbb{M}$ が選ばれたという条件の下で，未知パラメータを一定の確率で含む信頼区間を求めるべきだという考え方も存在する。そのような考えに則るならば，$\hat{\mathbb{M}}=\mathbb{M}$ を条件とする条件付確率が $1-\alpha$ になるような信頼区間を求めることが望ましい。しかし，この場合にも，未知パラメータについて一様に妥当な信頼区間を構築することは困難である（Leeb and Pötscher 2006）。

2.3 オラクル性を満たす正則化法

Tibshirani（1996）によって Lasso が提案されて以降，特に候補となる説明変数の数が多いときには，AIC や BIC などの情報量規準に基づく方法にとって代わって，正則化法が変数選択の方法としてよく用いられている。正則化法とは，推定値の取りうる値に制約をかけながら，パラメータの推定を行う方法である。Lasso であれば，次のような最小化問題を解くことで推定量を得る。

$$\min_{b\in\mathbb{R}^p}\sum_{i=1}^{n}(y_i-x_i'b)^2+\lambda\|b\|_1$$

ただし，$\|\bullet\|_1$ は L_1 ノルムを表す。λ はデータの分析者によって決定されるチューニングパラメータであり，この値を大きくするほど，推定値は絶対値で 0 に近い値をとる。

Lasso の利点は，変数選択と推定が同時に行える点にある。チューニングパラメータをうまく選んでやれば，いくつかの要素がぴったり 0 と推定されるためである。対応する係数が 0 と推定されるということは，その変数がモデルから除外されていることを意味する。また，候補となる説明変数の数が多いときには，情報量規準に基づいて変数選択を行うのは困難だが，Lasso ならば $p>n$ の場合でも低い計算負荷で変数選択が行える。一方弱点としては，Lasso はバイアスを持った推定量であり，漸近分布は正規分布とはならない。そのため，特定のパラメータの統計的推測が目的であるならば，Lasso は適していない。また，Lasso による変数選択は，一般には一致性を満たさない。

Fan and Li（2001）は，変数選択と推定が同時に行えるという正則化の利点を活かしつつ，統計的推測も行えるような方法を提案している。彼らの手法は次のような最小化問題を解く

ことで推定量を得る。

$$\min_b \sum_{i=1}^{n}(y_i - x_i'b)^2 + \sum_{j=1}^{p} p_\lambda(|b_j|)$$

ただし，

$$p_\lambda(|b_j|) = \begin{cases} \lambda|b_j| & |b_j| \leq \lambda \text{ のとき} \\ -\dfrac{|b_j|^2 - 2a\lambda|b_j| + \lambda^2}{2(a-1)} & \lambda < |b_j| \leq a\lambda \text{ のとき} \\ \dfrac{(a+1)\lambda^2}{2} & a\lambda < |b_j| \text{ のとき} \end{cases}$$

であり，$a>2$ もチューニングパラメータである。この正則化法は，SCAD（smoothly clipped absolute deviation）と呼ばれる。

SCAD が満たす重要な性質にオラクル性がある。真のパラメータベクトル β のいくつかの要素が 0 であるとする。一般性を失うことなく，最初の q 個の要素が非 0，残りの $p-q$ 個の要素が 0 であるとし，$\beta=(\beta_1', \beta_2')'=(\beta_1', 0')'$ と書く。それぞれ部分ベクトルに対応する SCAD 推定量を $(\hat{\beta}_1', \hat{\beta}_2')'$ とすると，λ が一定のレートで 0 に近づくとき

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, \Sigma), \quad P(\hat{\beta}_2 = 0) \to 1$$

を満たす。ただし，漸近共分散行列 Σ は，$\beta_2=0$ として β_1 のみを推定したときの OLS 推定量の漸近共分散行列と一致する。つまり，あらかじめ $\beta_2=0$ であることを知っている場合と同じ精度でパラメータを推定でき，このような性質はオラクル性と呼ばれる。また，β の 0 の要素はすべて 0 と推定されるので，これは変数選択の一致性が満たされることを意味する。これらの性質をみたす正則化法としては，SCAD の他にも Adaptive Lasso（Zou 2006）などがある。

SCAD は一見望ましい推定量のようだが，2.2 節で議論した手法と同様の問題を抱えていることが知られている（Leeb and Pötscher 2008b）。実際，SCAD 推定量の性質は，一致性を満たす方法で変数選択を行い，選ばれた変数だけを使って推定した OLS 推定量の性質と，非常によく似ている。そのため，SCAD 推定量を基に構築された信頼区間は，真のパラメータについて一様に妥当な信頼区間とはならない。また，著者の知る限りにおいては，オラクル性を満たすためには，候補となる説明変数の数はサンプルサイズよりも小さくなければならない。最近の高次元データ分析の文脈においては，これはかなり制約的な条件である。これらの理由から，オラクル性を満たす正則化法の研究は，現在では下火となっている。

3　変数選択後の妥当な統計的推測

3.1　統計モデルに対する2つの考え方

前節までの問題を踏まえ，以下ではデータ生成過程について一様に妥当な信頼区間の構築方法について考察する。また，候補となっている説明変数の数がサンプルサイズよりも大きいケースを許容する。このような設定の下で取られるアプローチは，何を興味のあるパラメータであると考えるかによって変わってくる。

パラメータに関するひとつめの考え方は，これまでの議論で暗に採用してきた考え方である。もう一度，2.1節の例を振り返る。そこでは，あるβの下，データは(1)によって生成され，興味のあるパラメータは(1)におけるパラメータβ_1であった。モデル$\mathbb{M}_1$はモデル$\mathbb{M}_2$における$\beta_2=0$のケースに対応すると考える。そのため，$\beta_2\neq0$であるときに$\mathbb{M}_1$を用いると，誤ったモデルを用いていることになり，$\hat{\beta_1}(\mathbb{M}_1)$には欠落変数バイアスが生じる。

ふたつめの考え方は，異なるモデルのパラメータにはそれぞれの解釈が与えられ，それぞれに興味のあるパラメータになりうるというものである。例えば，モデル$\mathbb{M}_1$をx_{i1}に基づくy_iの線形予測モデル，モデル$\mathbb{M}_2$をx_{i1}とx_{i2}に基づくy_iの線形予測モデルと考えれば，それぞれのモデルにおけるx_{i1}の係数には別の解釈が与えられ，それぞれに興味の対象となりうる。このような考え方に立つならば，そもそも(1)がデータの生成過程を正しく表している必要はないし，使用するモデルが正しく定式化されている必要もない。

このようなパラメータに対する考え方の違いは，構造的（structural）な分析と記述的（descriptive）な分析のどちらに興味があるかの違いである。構造的な分析，つまり，データの生成過程の分析に興味があるのであれば，正しいモデルが存在すると想定して，そのモデルの構造パラメータに関する統計的推測を考えるのが自然である。一方，記述的な分析，つまり，観測されるデータからわかる事実の分析のみに興味があるのであれば，便宜的に用いるモデルにおけるパラメータがわかればそれで十分であり，そのモデルがデータの生成過程（構造モデル）を表している必要はない。

スパース性の仮定の下，構造的な分析と親和的な統計的推測の方法がいくつか提案されている。スパース性とは，βの要素の数はサンプルサイズより大きくても構わないが，βに含まれる非0の要素の数はサンプルサイズよりずっと小さいという性質のことである。統計学のリテラチャーでは，例えばdebiased Lassoと呼ばれる方法が提案されている（Zhang and Zhang 2014; van de Geer et al. 2014など）。2.3節で述べたように，Lassoはバイアスを持ちそれ自体を統計的推測に用いることは困難だが，Lassoのバイアスを修正して，漸近的に正規分布に従うようにしたものがdebiased Lassoである。また，計量経済学のリテラチャーでは，Belloni et al.（2014）やChernozhukov et al.（2018）などが異なる手法を提案している。

しかし，これらの方法については，本稿ではこれ以上は論じない。

3.2　Berk et al.（2013）による PoSI 信頼区間

本節では，記述的な分析手法として，Berk et al.（2013）によって提案された PoSI（Post-Selection Inference）信頼区間を紹介をする。彼らは，変数選択の結果ランダムに選ばれたモデルのパラメータの信頼区間の構築方法を提案している。

設定は以下のとおりである。$y=(y_1, \ldots y_n)'$ とし，$y\sim N(\mu, \sigma^2 I)$ を満たすとする。ただし，I は n 次の単位行列である。分散 σ^2 については，既知もしくは，ある r について $\hat{\sigma}^2\sim\sigma^2\chi^2_r/r$ を満たす推定量 $\hat{\sigma}^2$ が存在することを仮定する。ただし，χ^2_r は自由度 r の χ^2 分布を表す。この推定量は特定のモデルには依存しないものとする。説明変数の行列 $X(n\times p)$ は必ずしもフルランクである必要はなく，$\mathrm{rank}(X)=d$ とする。

線形モデル $\mathbb{M}$ の説明変数の行列を $X_{\mathbb{M}}$ とし，そのモデルの説明変数の数を $m(\mathbb{M})$ とする。また，変数選択の結果実現しうるすべてのモデルの集合を $\mathcal{M}$ で表す。2.1 節の例では，モデルは 2 個しか考えていないので，$\mathcal{M}=\{\mathbb{M}_1, \mathbb{M}_2\}$ である。モデルの数は最大で 2^p 個であり，ネストしているモデルしか考えないなら p 個になる。変数選択後に各モデルは OLS で推定されるため，$\mathrm{rank}(X_{\mathbb{M}})=m(\mathbb{M})\leq d$ を満たすモデルのみを考察の対象とする。モデル $\mathbb{M}$ における興味のあるパラメータは

$$\beta(\mathbb{M})=(X'_{\mathbb{M}}X_{\mathbb{M}})^{-1}X_{\mathbb{M}}\mu=\arg\min_b\|\mu-X_{\mathbb{M}}b\|^2$$

であるとする。つまり，$X_{\mathbb{M}}$ による μ の最良線形近似を得ることが目的である。ここでのポイントは，y の分布はあくまでも μ によって決定されており，線形モデルは y の生成過程を表してはいないというところである。

変数選択が行われておらず，最初からモデル $\mathbb{M}$ を分析に用いているならば，$\beta(\mathbb{M})$ の各成分の信頼区間を求めることは容易である。OLS 推定量 $\hat{\beta}(\mathbb{M})$ の第 j 成分について，$\hat{\beta}_j(\mathbb{M})\sim N(\beta_j(\mathbb{M}), \sigma^2/\|X_{j\cdot\mathbb{M}}\|^2)$ が成り立つので，ある定数 K について，$\beta_j(\mathbb{M})$ の信頼区間を

$$\mathrm{CI}_j^K(\mathbb{M})=\left[\hat{\beta}_j(\mathbb{M})-K\frac{\hat{\sigma}}{\|X_{j\cdot\mathbb{M}}\|},\ \hat{\beta}_j(\mathbb{M})+K\frac{\hat{\sigma}}{\|X_{j\cdot\mathbb{M}}\|}\right] \qquad (6)$$

によって求めることにする。ただし，$X_{j\cdot\mathbb{M}}$ は $X_{\mathbb{M}}$ の第 j 列をその他の列に回帰した残差のベクトルを表す。t 統計量

$$t_j(\mathbb{M})=\frac{\hat{\beta}_j(\mathbb{M})-\beta_j(\mathbb{M})}{\hat{\sigma}/\|X_{j\cdot\mathbb{M}}\|}$$

について $|t_j(\mathbb{M})|\leq K$ であることと，$\beta_j(\mathbb{M})\in\mathrm{CI}_j^K(\mathbb{M})$ であることは同値なので，K として自由度 r の t 分布の上側 $\alpha/2$ 点を選ぶと

$$P(\beta_j(\mathbb{M})\in\mathrm{CI}_j^K(\mathbb{M}))=1-\alpha$$

を満たす。ところが，変数選択の結果得られたモデルを $\hat{\mathbb{M}}$ とすると，同じ K について，一般に $\mathrm{CI}_j^K(\hat{\mathbb{M}})$ は $\beta_j(\hat{\mathbb{M}})$ に対する妥当な信頼区間とはならない。

Berk et al.（2013）の目的は，どのような方法で選択された $\hat{\mathbb{M}}$ についても

$$P(\beta_j(\hat{\mathbb{M}}) \in \mathrm{CI}_j^K(\hat{\mathbb{M}}),\ j=1,\ldots,m(\hat{\mathbb{M}})) \geq 1-\alpha \qquad (7)$$

を満たすような定数 K を見つけることである。この定数を PoSI 定数と呼び，PoSI 定数を用いた（6）の信頼区間を PoSI 信頼区間と呼ぶ。変数選択の方法は問わないので，恣意的な方法で選択されたモデルのパラメータに対しても，妥当な信頼区間を構築することができることが特徴である。PoSI 信頼区間の意味するところを直感的に述べると，次のようになる。仮に y を100回，$N(\mu, \sigma^2 I)$ から抽出すると，標本ごとに異なるモデルが選択される可能性があるが，彼らの方法で信頼区間を構築すれば，95回以上は興味のあるパラメータを含むものになっている。ただし，興味のあるパラメータそのものも標本に依存して変化するので，100個の信頼区間は同じパラメータに対する信頼区間には必ずしもなっていない。

上記の説明からもわかるとおり，Berk et al.（2013）の問題設定は 2 点において従来のものとは異なっている。ひとつは興味のあるパラメータがモデルごとに異なっている点，もうひとつは $\hat{\mathbb{M}}$ がランダムに変動するため，興味のあるパラメータ $\hat{\beta}(\hat{\mathbb{M}})$ も確率変数になっている点である。通常の信頼区間の構築においては，パラメータは固定された定数であると考えるが，PoSI の設定ではパラメータと区間がともにランダムに変動する。

PoSI 定数の求め方のアイデアは以下のとおりである。信頼区間と t 統計量の関係より，（7）が満たされるためには，同じ定数 K について

$$P(\max_{1\leq j\leq m(\hat{\mathbb{M}})} |t_j(\hat{\mathbb{M}})| \leq K) \geq 1-\alpha$$

が成り立てばよい。また，明らかに

$$\max_{1\leq j\leq m(\hat{\mathbb{M}})} |t_j(\hat{\mathbb{M}})| \leq \max_{\mathbb{M}\in\mathcal{M}} \max_{1\leq j\leq m(\mathbb{M})} |t_j(\mathbb{M})|$$

が成立するので，上式の右辺を確率 $1-\alpha$ でバウンドできるような定数 K を求めればよい。t 統計量の分布は未知母数 μ と σ^2 に依存しないので，PoSI 定数 K を用いれば，いかなる μ と σ^2 についても（7）が成立し，従来の手法で問題となっていた一様性の問題が解決される。

PoSI 定数の求め方について，さらに詳細な議論はしないが，特殊なケースを除いて解析的に求めることはできず，コンピュータを用いて数値的にしか求められない。また，説明変数の数が多いときには計算負荷が高く，著者らは $\mathrm{rank}(X) > 20$ の場合には求めることが難しいと述べている。

3.3　正確な被覆確率をもつ信頼区間

Berk et al.（2013）の方法は，どのような方法で変数選択が行われても妥当な信頼区間が求められるという利点があるが，実際の被覆確率は意図した信頼水準よりも大きくなりがち

で，信頼区間も広くなる傾向にある。変数選択の方法を特定の方法に限定すれば，被覆確率が厳密に $1-\alpha$ になるような信頼区間を構築でき，信頼区間を狭くできる可能性がある。

Lee et al.（2016）と Tibshirani et al.（2016）は，Lasso などの特定の方法で変数選択が行われた場合の変数選択後の信頼区間の求め方を考察している。彼らは，あるモデル $\mathbb{M}$ が選択された後に，定数のベクトル $\eta\in\mathbb{R}^n$ について，興味のあるパラメータを $\eta'\mu$ と定義し，次のような性質を満たす信頼区間 CI を求めることを提案している。

$$P(\eta'\mu\in \mathrm{CI}\,|\,\hat{\mathbb{M}}=\mathbb{M})=1-\alpha$$

特に，$e_j\in\mathbb{R}^{m(\mathbb{M})}$ を第 j 成分のみが 1 でその他の成分が 0 のベクトル，$\eta=X_{\mathbb{M}}(X'_{\mathbb{M}}X_{\mathbb{M}})^{-1}e_j$ とすれば，$\eta'\mu=\beta_j(\mathbb{M})$ となる。$y\sim N(\mu, \sigma^2 I)$ であるとして，σ^2 は既知であるとしておく。

方法の説明に入る前に，Berk et al.（2013）の信頼区間との違いについて触れておく。Berk et al.（2013）の PoSI 信頼区間とは異なり，ここではある特定のモデル $\mathbb{M}$ が選ばれたケースのみを考察の対象としている。この信頼区間の意味するところを直感的に述べると，次のようになる。y を $N(\mu, \sigma^2 I)$ から何度も抽出し，毎回 Lasso などの決められた方法で変数選択をする。そのうちモデル $\mathbb{M}$ が選ばれたケースのみを100回取り出し，それぞれの場合で信頼区間を求めるとする。すると，100個の信頼区間のうちおよそ95個は $\eta'\mu$ を含んでいる。

Lee et al.（2016）と Tibshirani et al.（2016）は，y の正規性と，特定の変数選択法の性質を利用すれば，事象 $\{\hat{\mathbb{M}}=\mathbb{M}\}$ を条件とした下での $\eta'y$ の条件付分布がシンプルな形で特徴づけられることを利用して，$\eta'\mu$ の信頼区間を構築する。以下では，Lasso を用いて変数選択をすることを想定するが，LAR（Efron et al. 2004）や変数増加法（forward stepwise）でも同様の結果は得られる。X を説明変数の行列とするときの Lasso 推定量を $\hat{\beta}$ とすると，Lasso によって選択される変数の集合は，$\hat{\mathbb{M}}=\{j\in\{1, \ldots, p\}:\hat{\beta}_j\neq 0\}$ である。つまり，最初にすべての説明変数を用いて Lasso 推定を行った結果，対応する係数が 0 にならなかった変数のみを取り出す。

Lasso による変数選択の特徴は，それぞれのモデルに対応する y の集合が，多面体の集合によって表されることである。より正確には次のようになる。Lasso 推定量 $\hat{\beta}$ によって選択されたモデルを $\hat{\mathbb{M}}$，$\hat{\beta}$ の非 0 成分の符号を表すベクトルを $\hat{s}\in\{-1, 1\}^{m(\hat{\mathbb{M}})}$ とする。例えば，$\hat{\beta}=(1.2, 0, 0, -2.5, 4.1)'$ であれば，$\hat{\mathbb{M}}=\{1, 4, 5\}$，$\hat{s}=(1, -1, 1)'$ である。このとき，ある行列 A_s とベクトル b_s が存在して

$$\{y\in\mathbb{R}^n:\hat{\mathbb{M}}=\mathbb{M}, \hat{s}=s\}=\{y\in\mathbb{R}^n:A_s y\leq b_s\}$$

となることが Lasso の最小化問題の KKT 条件から示される[2)]。つまり，ある特定のモデルと符号ベクトルを実現するような y の集合は，$\{Ay\leq b\}$ という集合で表される。

さらに，y を η が張る空間に射影した射影残差を z とすると，A_s と b_s に依存する関数 $(\mathcal{V}_s^-, \mathcal{V}_s^+, \mathcal{V}_s^0)$ が存在して

$$\{A_s y \leq b_s\} = \{\mathcal{V}_s^-(z) \leq \eta' y \leq \mathcal{V}_s^+(z),\ \mathcal{V}_s^0(z) \geq 0\}$$

と表現できる。また，y の正規性から $\eta' y \sim N(\eta'\mu,\ \sigma^2\|\eta\|^2)$ で，z とは独立となる。これらをまとめると，$\{\hat{\mathbb{M}} = \mathbb{M},\ \hat{s} = s\}$ を条件とすると，$\eta' y$ は $N(\eta'\mu,\ \sigma^2\|\eta\|^2)$ を下限 $\mathcal{V}_s^-(z)$ と上限 $\mathcal{V}_s^+(z)$ で切断した切断正規分布に従うことがわかる[3)]。よって，集合 S 上で切断された $N(\mu,\ \sigma^2)$ の分布関数を F^S_{μ,σ^2} とすると

$$F^{[\mathcal{V}_s^-(z),\mathcal{V}_s^+(z)]}_{\eta'\mu,\ \sigma^2\|\eta\|^2}(\eta' y) \,|\, \{\hat{\mathbb{M}} = \mathbb{M},\ \hat{s} = s\} \sim \mathrm{Unif}[0,\ 1]$$

が成立する[4)]。

求める信頼区間の構築のためには，$\{\hat{\mathbb{M}} = \mathbb{M}\}$ を条件とする条件付分布が既知であるような統計量が必要である。ここで，$\{\hat{\mathbb{M}} = \mathbb{M},\ \hat{s} = s\}$ の起こりうる符号 s に関する和集合を取ることで，$\{\hat{\mathbb{M}} = \mathbb{M}\} = \cup_s \{\hat{\mathbb{M}} = \mathbb{M},\ \hat{s} = s\}$ と表される。そのため，$\{\hat{\mathbb{M}} = \mathbb{M}\}$ に条件付けることは，$\cup_s \{A_s y \leq b_s\}$ で条件付けるのと同じであり，

$$F^{\cup_s[\mathcal{V}_s^-(z),\mathcal{V}_s^+(z)]}_{\eta'\mu,\ \sigma^2\|\eta\|^2}(\eta' y) \,|\, \{\hat{\mathbb{M}} = \mathbb{M}\} \sim \mathrm{Unif}[0,\ 1]$$

が成り立つことも示される。$\eta' y$ の信頼水準 $1-\alpha$ の信頼区間を求めるには，区間の下限 L と上限 U をそれぞれ

$$F^{\cup_s[\mathcal{V}_s^-(z),\mathcal{V}_s^+(z)]}_{L,\ \sigma^2\|\eta\|^2}(\eta' y) = 1 - \frac{\alpha}{2}, \quad F^{\cup_s[\mathcal{V}_s^-(z),\mathcal{V}_s^+(z)]}_{U,\ \sigma^2\|\eta\|^2}(\eta' y) = \frac{\alpha}{2}$$

を満たすように選べばよい。信頼区間の構築に用いる統計量の条件付分布は μ に依存しないため，この信頼区間は μ について一様に妥当な信頼区間となっている。

4 おわりに

本稿では，変数選択後の信頼区間の構築方法について，これまで慣例的に用いられてきた方法の問題点を指摘するとともに，Berk et al.（2013）などによって提案された新しい信頼区間の構築方法を紹介した。これらの新しい方法は，変数選択の結果選ばれたモデルに依存して信頼区間を構築するパラメータを決めるのが特徴で，p 値ハッキングのような問題に対処する方法のひとつとなりうるかもしれない。

もちろん，これらの方法を用いればすべての問題が解決するわけではない。技術的な側面について言えば，これらの論文では正規性などいくつかの強い仮定が置かれている。ただし，この点については，Bachoc et al.（2020）や Tibshirani et al.（2018）が，より広いクラスの分布について一様に妥当な信頼区間の構築方法を提案しており，今後さらに優れた方法が提案される可能性もある。

より本質的な問題は，そもそもこれらの論文で考察されているパラメータが，データの分析者にとって興味のあるパラメータとなりうるかどうかであろう。経済学の実証研究の多くは，当然のことながら経済理論と結びついており，構造的な分析が主たる目的となることが

多いと考えられる。そのようなときに，記述的な分析法を用いることにどれだけの意味があるのかはわからない。しかし一方で，データ生成過程を完全に記述できるモデルが存在すると考えるのはあまりに楽観的であり，どうせ正しいモデルを得ることなど不可能であれば，近似モデルを用いた記述的な分析にも一定の役割はあるかもしれない。

注

本研究は，JSPS 科研費 19H01473 による助成を受けたものである。

1）$\beta_2=0$ であるとき，$\mathbb{M}_1$ と $\mathbb{M}_2$ はともに正しいモデルである。2つのモデルの違いは，$\mathbb{M}_1$ では β_2 を既知の値（$\beta_2=0$）としているのに対し，$\mathbb{M}_2$ では β_2 を未知パラメータとして扱っている点にある。$\mathbb{M}_0$ は単に正しいモデルではなく，正しく定式化されたモデルの中で最も未知パラメータが少ないものを指す。

2）A_s と b_s はモデル $\mathbb{M}$ にも依存するが，表記の単純化のため記号は省略する。

3）実際には z は確率変数なので，z をある値 z_0 で条件付けたときに，下限 $\mathcal{V}_s^-(z_0)$ と上限 $\mathcal{V}_s^+(z_0)$ の切断正規分布に従うというのが正確であるが，ここでの議論では特に問題とはならない。

4）任意の確率変数 X について，その分布関数を $F(x)$ とすると，$F(X)$ は $[0, 1]$ 区間の一様分布に従う。

参 考 文 献

Bachoc, F., D. Preinerstorfer, and L. Steinberger (2020). “Uniformly valid confidence intervals post-model-selection.” *Annals of Statistics 48*(1), 440-463.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). “Inference on treatment effects after selection among high-dimensional controls.” *Review of Economic Studies 81*(2), 608-650.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). “Valid post-selection inference.” *Annals of Statistics 41*(2), 802-837.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). “Double/debiased machine learning for treatment and structural parameters.” *Econometrics Journal 21*, C1-C68.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). “Least angle regression.” *Annals of statistics 32*(2), 407-499.

Fan, J. and R. Li (2001). “Variable selection via nonconcave penalized likelihood and its oracle properties.” *Journal of the American Statistical Association 96*(456), 1348-1360.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). “Exact post-selection inference, with application to the lasso.” *Annals of Statistics 44*(3), 907-927.

Leeb, H. and B. M. Pötscher (2005). “Model selection and inference: Facts and fiction.” *Econometric Theory 21*(1), 21-59.

Leeb, H. and B. M. Pötscher (2006). “Can one estimate the conditional distribution of post-model-selection estimators?” *Annals of Statistics 34*(5), 2554-2591.

Leeb, H. and B. M. Pötscher (2008a). “Can one estimate the unconditional distribution of post-model

-selection estimators?" *Econometric Theory 24*(2), 338–376.

Leeb, H. and B. M. Pötscher (2008b). "Sparse estimators and the oracle property, or the return of hodges' estimator." *Journal of Econometrics 142*(1), 201–211.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B 58*(1), 267–288.

Tibshirani, R. J., A. Rinaldo, R. Tibshirani, and L. Wasserman (2018). "Uniform asymptotic inference and the bootstrap after model selection." *Annals of Statistics 46*(3), 1255–1287.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016). "Exact post-selection inference for sequential regression procedures." *Journal of the American Statistical Association 111*(514), 600–620.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models." *Annals of Statistics 42*(3), 1166–1202.

Zhang, C.-H. and S. S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models." *Journal of the Royal Statistical Society. Series B 76*(1), 217–242.

Zou, H. (2006). "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association 101*(476), 1418–1429.