Lasso 回帰における変数選択と推定の不安定性に対するアルゴリズムの提案

Approaches for Algorithms in Lasso Regression: Addressing Variable Selection and Estimation Instability

井出　草平/Sohei IDE[1]
[1] 多摩大学情報社会学研究所 客員教授 [a]

［Abstract］This paper conducts simulations on the stability of variable selection and estimation in Lasso regression (Least Absolute Shrinkage and Selection Operator) and proposes new methodologies. Lasso regression is widely used in biology, medicine, social sciences, and machine learning. It not only automatically selects the most relevant explanatory variables but also computes their corresponding estimates with a high degree of efficiency. Although Lasso regression is a powerful statistical analysis method, its use of random numbers during computation causes variability in the selected variables and their estimates. Reproducibility is a critical factor in science, but the inherent variability due to random numbers in Lasso regression poses a significant challenge to reproducibility. This paper evaluates various methods proposed for Lasso regression by conducting repeated simulations with different random seeds on actual data. Specifically, it assesses the performance of holdout, K-fold cross-validation, repeated cross-validation, Monte Carlo cross-validation, nested cross-validation, leave-one-out cross-validation, Adaptive Lasso, stability selection, randomized stability selection, Forward Stability, and Model Path Selection through simulations. Building on methods proposed in previous studies, this research proposes four new algorithms to ensure reproducibility in statistical modeling. These methods differ from previous approaches and aim to guarantee stable statistical model estimation, thereby contributing to improved reproducibility across various research fields.

［キーワード］
Lasso 回帰、交差検証、クロスバリデーション、安定性選択

1. はじめに

本論文では、Lasso 回帰（Least Absolute Shrinkage and Selection Operator）における変数選択の安定性と推定量の安定性に関するシミュレーションを行い、新しい方法論を提案する。Lasso 回帰は Tibshirani (1996)の提案した統計手法であり、現在では生物学、医学、社会科学、機械学習の分野で広く利用されている。そのため、日本語でも Lasso 回帰の解説書やウェブ上の情報は数多く存在する(馬場, 2018)。日本語文献のほとんどは機械学習の分野に関するものである。また、英語文献においても大半が機械学習分野のものであり、統計学と機械学習の関心の違いから、Lasso 回帰の特徴や欠点に関する考察や欠点を克服するための統計技法の開発は十分とは言い難い。そこで本論文では、応用統計学の立場から Lasso 回帰の方法論についてシミュレーションを用いて性能評価を行い、新しい方法論を提案する。これにより、Lasso 回帰の運用は社会科学や医学などの研究分野に影響を与えるだけでなく、機械学習分野においても正確で安定した推定方法を提供できるであろう。

Lasso 回帰は、回帰分析の一種であり、特に高次元データの次元削減や変数選択に使用される手法である。データセットに多くの変数がある場合、従来は相関係数やカイ二乗検定などを用いて一つ一つ関連を調べ、その後に回帰分析の独立変数として投入する手順が一般的である。しかし、この方法は手作業であり、最適解が選ばれている確証はない。Lasso 回帰は、従属変数と関連のない独立変数を大量に投入しても、アルゴリズムが自動的に関連のある変数を選択し、独立変数の係数を計算する。このため、Lasso 回帰は統計解析において非常に強力な分析法と考えられている。特に、多くの変数の中から自動的に説明力のある変数を選択するという特徴は機械学習との相性がよく、機械学習において欠かせない解析方法の一つとなっている。

Lasso 回帰には大きく 3 つの特徴がある。第 1 に、次元削減である。不要な変数を自動的に除去することで、モデルを簡潔にする。第 2 に、過学習防止である。パラメータの絶対値の和をペナルティ項として追加することで、過学習を防ぐ。第 3 に、解釈性の向上である。重要な変数のみを残すため、モデルの解釈が容易になる。

Lasso 回帰は次のような数式で表現される。

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}\right)^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

[a] sohei@ni.tama.ac.jp

nはサンプル数、y_iは目的変数のi番目の観測値、$x_i^T \hat{\beta}$ は説明変数ベクトルχ_iと回帰係数ベクトル$\hat{\beta}$の内積である。$(y_i - x_i^T \hat{\beta})^2$ は各データポイントにおける予測誤差の二乗を表し、モデルの予測誤差の二乗和を最小化する。

右辺右側のλは正則化パラメータでありペナルティの重みを決定する。mは説明変数の数、$|\widehat{\beta_j}|$は回帰係数の絶対値を意味する。Lasso 回帰は回帰係数の絶対値の和に対してペナルティをかけることで、独立変数の中の変数の係数をゼロにする。係数がゼロになった変数はモデルに組み込まれず、説明力のある変数のみが、モデルに組み込まれることになる。

Lasso 回帰は数理的に洗練された方法論であるが、λ を計算する際に乱数を用いるため、大きな問題を抱えている。計算機で異なる乱数シードを使用すると異なる λ が選ばれ、その結果として異なる変数の組み合わせが選択される。また、同じ変数が選択されても異なる推定値が計算される。これは、Lasso 回帰が乱数を用いる以上避けられない問題である。

変数選択や推定量が計算のたびに異なるという点は大きな問題である。なぜなら、変数の処理を正しく行い、正確に計算しても、毎回異なる変数が選ばれ、異なる推定量が提示されるからである。例えば、論文で数値を報告しても、その数値は計算時の値に過ぎず、再計算すれば別の変数が選択され、異なる推定量が得られる可能性がある。このように、Lasso 回帰は非常に強力な統計手法であるが、再現性の点では非常に不安定であるという欠点が存在する。

Lasso 回帰にはさまざまな交差検証の方法が提案されており、本論文のテーマである変数選択と推定量の安定性を目的とした安定性選択（Stability Selection）やそこから派生した手法も存在する。本論文では、実データを用いたシミュレーションを通じて、先行研究で提案されている方法論の変数選択について評価する。さらに、シミュレーション結果に基づき、変数選択と推定値の安定性を向上させるアルゴリズムを提案する。

2. 方法

2.1. シミュレーションに用いるデータ

シミュレーションに使用するデータは StataCorp LLC が提供している車の機能の価格に関するデータ”auto.dta”である。Stata のウェブページからダウンロードができる(http://www.stata-press.com/data/r9/auto.dta)。データの説明は表-1にまとめた。

表- 1 シミュレーションに使用する変数の項目

変数名	説明	変数型
make	車のメーカーとモデル名	文字列型
price	車の価格（ドル）	数値型
mpg	ガソリン 1 ガロンあたりの走行マイル数	数値型
rep78	1978 年の修理記録（1-5 のスケール、5 が最高）	順序型
headroom	ヘッドルーム（フィート）	数値型
trunk	トランクの容量（立方フィート）	数値型
weight	車の重量（ポンド）	数値型
length	車の長さ（インチ）	数値型
turn	U ターンのための最小回転半径（フィート）	数値型
displacement	エンジンの排気量（立方インチ）	数値型
gear_ratio	ギア比	数値型
foreign	車の原産国（0 = 国内、1 = 輸入）	因子型

ケースには車の車種が入っており”Honda Accord”の場合は”price”が”5,799”、”mpg”が”25”というようにデータが格納されており、全部で 69 車種のデータが含まれている。したがって、”auto.dta”データはケース数 69 のデータである。

2.2. シミュレーションに用いる交差検証の方法

Lasso 回帰ではモデルの適合性と一般化性能を評価するために交差検証(Cross-Validation)が用いられる。交差検証とは、データを訓練データと検証データに分けてモデルの性能を評価する手法である。これにより、モデルが新しいデータに対してどの程度一般化できるかを検証することができる。交差検証は Stone(1974)によって提唱された方法であり、現在では様々な方法論が提案されている。

交差検証の方法は、網羅的交差検証と非網羅的交差検証の 2 つに大別されている(Arlot & Celisse, 2010)。網羅的交差検証は、元のサンプルを訓練セットと検証セットに分割するすべての可能な方法を試し、モデルを学習およびテストする手法である。一方、非網羅的交差検証は、すべての組み合わせを検証する leave-p-out 交差検証の近似手法であり、計算コストが比較的低いのが特徴である。

本論文では、非網羅的交差検証としてホールドアウト法、K 分割交差検証、反復交差検証、モンテカルロ交差検証、入れ子交差検証を検討する。網羅的交差検証としては一つ抜き交差検証を取り上げる。また、交差検証の方法とは異なるアプローチではあるが、Adaptive Lasso も本論文のテーマに関連があるため、併せて取り上げる。

2.2.1. ホールドアウト法

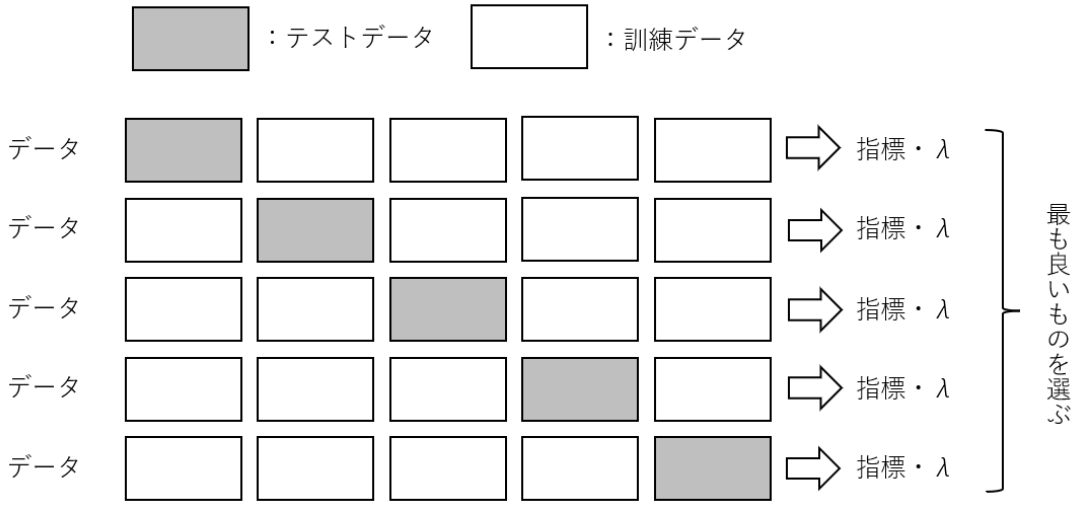
ホールドアウト法は交差検証の最初に考案された方法で、データセットを一度だけ訓練データと検証データに分割する手法である(Bishop, 2016)。具体的には、データセットの一部（例えば 70%）を訓練データとして使用し、残りのデータ（30%）を検証データとして用いる。この方法は計算コストが低く、実装も容易である点が特徴である。

Lasso 回帰においてホールドアウト法を使用する場合、まずデータを訓練データと検証データに分割する。次に、訓練データを用いてモデルを構築し、λ の最適な値を選択する。最適な λ が決定した後、その値を用いて最終的なモデルを訓練し、検証データでモデルの性能を評価する。

ホールドアウト法による交差検証の結果はデータの分割方法に依存するため、データセットの構成によっては結果がばらつくことがある。特にサンプルサイズがそれほど多くない場合、訓練データと検証データに分けることは、結果のばらつきに大きく影響することがある。そのため、ホールドアウト法が Lasso 回帰で使われることはほとんどない。

2.2.2. K 分割交差検証

K 分割交差検証は、Geisser(1975)によって提案されたもので、データセットを k 個のフォールド（部分集合）に分割し、モデルの性能を評価するための手法である。概念図を図-１に示した。データを k 個の部分集合に分割し、各部分集合を検証データ、残りを訓練データとして使用する。このプロセスを k 回繰り返し、各 λ に対する平均二乗誤差（MSE）などの評価指標を計算し、最も低い評価指標を持つ λ を選択する 。計算コストが他の交差検証より比較的低いことが利点として挙げられる。欠点は 各フォールドの分割が異なるため、得られる評価指標にばらつきが生じる可能性があるである。一般的に k = 10、つまり 10-fold 交差検証が選ばれることが多い(McLachlan et al., 2004)。機械学習の分野をはじめとして、交差検証の方法としては最も選択されている方法である(He et al., 2023)。



図－１K 分割交差検証の概念図

2.2.3. 反復交差検証

反復交差検証（Repeated Cross-Validation）とは、データを複数回にわたり異なる分割パターンで訓練データと検証データに分割し、モデルの性能を評価する手法である(Kohavi, 1995)。具体的には、K 分割交差検証を複数回繰り返す。各反復において、データを再度ランダムに分割し、新たな分割パターンで K 分割交差検証を行う。反復を行うことで、K 分割交差検証を 1 回行った場合よりも、モデルが頑健になることが期待できる。特に、データの分割によるバイアスを減少させることが可能である。

文献には、反復回数は 5 回から 10 回の間が推奨されていることが多いが、その根拠は経験則に基づいているようだ。社会調査で使用するデータは多くても数千ケースであり、反復計算に必要な時間はそれほど長くない。しかし、機械学習やビッグデータの処理において反復計算が必要である場合、反復回数が増加すると計算時間も膨大になり、実際の運用が困難になる。計算時間と計算の安定性は、トレードオフの関係にある。

反復交差検証のシミュレーションでは、反復回数についても検証を行う。具体的には、機械学習やビッグデータを扱うわけではないが、多くても数千件程度のデータセットの場合に望ましい反復回数をシミュレーションで確認する。結果は補足資料において示した。

2.2.4. モンテカルロ交差検証

モンテカルロ交差検証(Monte Carlo cross-validation, MCCV)、もしくは反復ランダムサブサンプリング検証(Repeated random sub-sampling validation)と呼ばれる方法は、訓練データと検証データをランダムに分割し、ホールドアウト法と同じく訓練データに基づいて検証データモデル評価を行うという手順を複数回(例えば 100 回)行う(Xu & Liang, 2001)。ランダム分割の数が無限大に近づくにつれて、p-個抜き交差検証の結果に近づく。反復交差検証とよく似た手順で行われるが、反復交差検証は K 分割交差検証を繰り返すのに比べ、モンテカルロ交差検証はホールドアウト法を繰り返すところに違いがある。

2.2.5. 入れ子交差検証

入れ子交差検証（Nested Cross-Validation）は、モデルの性能評価と λ の選定におけるバイアスを排除するための検証手法である(Stone, 1977)。外側の反復でモデルの性能を評価するデータセットと、λ の選定を行うデータセットを分けることに特徴がある。この手順を踏む理由は、一般的な K 分割交差検証の問題を克服するためである。

K 分割交差検証では、データセットを k 個のフォールドに分割し、それぞれのフォールドを一度ずつ検証データとして使用し、残りのフォールドを訓練データとして使用する。この方法では、λ の選定とモデルの性能評価が同じデータセットで行われるため、バイアスが生じる可能性がある。λ はモデルの性能評価によって選定され、モデルの性能評価には RMSE や MAE といった統計量を使用する。入れ子交差検証では、この選定と評価が同じデータセットで行われることによるバイアスを防ぐことが目的である。

入れ子交差検証では、各フォールドで得られた性能指標を平均し、異なる λ の値に対してこのプロセスを繰り返す。最も良い性能を示した λ の値が最適な λ として選ばれる。λ を選定した後は、K 分割交差検証と同じ手順で Lasso の推定が行われるため、仮に入れ子交差検証と K 分割交差検証で選定された λ が同じであれば、Lasso の推定値も同じになる。要するに、入れ子交差検証は λ の選定過程をより厳密にすることで、より良い λ の選定を行うための技法である。

2.2.6. 一つ抜き交差検証

一つ抜き交差検証（Leave-One-Out Cross-Validation, LOOCV）は、各データポイントを一度だけ検証データとして使用し、残りのデータを訓練データとして使用する交差検証である(Stone, 1977; Geisser, 1975; Molinaro et al., 2005)。具体的には、まずデータセットから 1 つのデータポイント（1 ケース分のデータ）を検証データとして抜き出し、残りを訓練データとする。次に、この訓練データを用いてモデルを訓練する。訓練されたモデルを、先に抜き出した 1 つのデータポイントを用いて評価し、このとき評価指標を計算する。この手順をデータセットの全データポイントについて繰り返す。最後に、全データポイントに対する評価指標を集計し、平均やその他の統計量を計算することで、モデルの性能を評価する。

一つ抜き交差検証の利点としては、各ステップで全てのデータを訓練に使用するため、モデルの評価にバイアスが少ない点が挙げられる。しかし、全データを用いて訓練されるために作成されたモデルはデータに非常によくフィットするものとなる反面、訓練データのノイズや偶然の変動まで学習してしまうことがあり、これにより

「過学習」が生じる可能性がある。過学習は、モデルが訓練データに対しては高い精度を示すが、新しいデータに対しては汎化性能が低下する現象である。さらに、一つ抜き交差検証は計算コストが高い点も欠点として挙げられる。データポイントの数だけモデルを訓練する必要があるため、大規模データセットでは計算コストが非常に高くなる。

2.2.7. Adaptive Lasso

Adaptive Lasso は交差検証の方法ではないが、同じ文脈で検討すべき方法なので、シミュレーションを行う(Zou, 2006)。通常の Lasso 回帰では、すべての変数に対して同じペナルティを課すが、Adaptive Lasso では変数ごとに異なるペナルティを課すことで変数選択の精度を向上させるところに特徴がある。Adaptive Lasso の具体的な手順は、まず通常の Lasso 回帰を用い、計算された係数の逆数をペナルティとして使用する。Lasso で選択されなかった変数には非常に大きな数値（例えば 10 の 6 乗）のペナルティを設定し、Adaptive Lasso の計算から除外する。

Adaptive Lasso はこの手順を踏むことで、重要な変数をより正確に選択し、真のモデルに近い結果を得やすくする。一方で、事前に行う初期推定の精度に依存するため、初期推定が不正確な場合、最終的な推定結果にも影響を与える可能性がある。

2.3. 安定性選択

交差検証を正確に行うアプローチとは違ったアプローチで安定的な変数選択を行うのが、安定性選択（Stability Selection）と呼ばれる技法である。この方法は Meinshausen と Bühlmann (2010)が提案し、Shah と Samworth(2013)が発展させたものである。この手法は、カットオフ値を設けて変数の選択を行うところに特徴がある。具体的には、サブサンプルを複数回（例えば 50 回や 100 回）抽出し、各サブサンプルに対して Lasso 回帰を実行して変数選択を行う。その結果を変数リストに保存し、各変数がサブサンプルから選ばれた頻度を計算して選択確率を求めるという方法である。

また、安定性選択はエラー制御を重視する方法であるが、重要な予測子が見逃されるリスクを減らすことを目的としたランダム化された安定性選択（Randomized Stability Selection）も提案されている。これらの方法は変数選択を安定的に行うことを目的としており、変数選択後の推定はこの方法の提案には含まれていない。

安定性選択やランダム化された安定性選択には、分析者が決めるべき数値が最低でも 3 つ存在する。1 つ目は変数を選択するカットオフ値である。カットオフ値は 0.75 を採用した(Afreixo et al., 2024)。2 つ目は選択確率の分布であり、より一般的な形状を持つ「R-concave」を採用した。3 つ目は PFER（Per-Family Error Rate）である。PFER は安定性選択において許容される誤選択（false positives）の上限を表し、1 が使用されることが多いため 1 を選択した。この場合、平均して 1 つ以下の誤選択が許容されることになる。安定性選択とランダム化された安定性選択を同じ条件で実施した。

安定性選択とランダム化された安定性選択には、結果がスパースになるという問題が指摘されている。要は選ばれる変数が少なく、モデルがスリムになりすぎるのである。この問題に対応するため、前進選択法(Forward Stability)とモデル経路選択(Model Path Selection)という方法が提案されている(Kissel & Mentch, 2024)。Kissel と Mentch らの方法は Shah と Samworth(2013) の発想を基にしており、安定性選択の一種と見なすことができる。

前進選択法は、データセットを複数のサブセットに分割し、各サブセットに対してモデルをフィッティングして変数選択を行う。選択頻度が一定の閾値を超えた変数を最終モデルに含める。この方法は安定した変数選択を実現し、複数の構造的に類似したモデルを生成し、モデル選択の不確実性を可視化できる点が優れている。一方で、計算コストが非常に高いという欠点がある。また、閾値の設定は分析者に依存するため、適切な選択が必要である。

モデル経路選択は、初期モデルを選び、一段階ずつ変数を追加し、各ステップで最適な変数を選択する。このプロセスを繰り返し、構築されたモデルパスを保存する。利点は安定した変数選択を実現する点で前進選択法と同様であるが、やはり計算コストが高い点が欠点である。シミュレーションではいずれも 3 階層モデルで推定を行った。

2.4. 変数選択のシミュレーション方法

シミュレーション方法は、この論文全体を通じて同じ手順で行う。Lasso 回帰が変数選択と推定値で異なる結果を示す原因は、λ を計算に使用する乱数にある。そこで、異なる乱数シードを設定して 100 回反復計算を行い、

その結果を比較する。反復回数は多いほど正確ではあるが、本論文で使用するデータを用いてLasso回帰の反復計算をしたところ、数十回程度で十分な結果を示しており、余裕をみて反復回数を100回とした。Lasso回帰の変数選択と推定値の安定性を評価するために、以下の2つの指標を算出した。

1) 独立変数として選ばれた変数の回数
2) Jaccard係数

Jaccard係数とは、集合の類似度を測るための統計的指標であり、1が完全一致、0が完全不一致である(Jaccard, 1901)。今回は、100回の反復計算の結果の一致度を評価することを目的としている。そのため、100個の結果を比較するために平均Jaccard係数を用いることとした。

計算にはR Ver. 4.4.0を使用した。また、Lassoと交差検証の計算にはglmnetパッケージVer.4.1-8、caretパッケージVer.6.0-94、nestedcパッケージVer.0.7.8を使用し、安定性選択にはstabsパッケージVer.0.6-4、ランダム化された安定性選択にはmonaLisaパッケージVer.1.10.0、シミュレーションの反復計算にはdoParallelパッケージVer.1.0.17を使用した。

3. 結果

3.1. 交差検証ごとの変数選択の安定性

本論文で焦点を当てているLasso回帰の変数選択の問題点は、計算のたびに全く異なった変数が選択され、その推定値も異なるという点である。この問題が生じる原因は、Lasso回帰のλを計算する際に使用する乱数にあり、その乱数によって結果が異なることにある。Lasso回帰のこの欠点を補うため、様々な交差検証の方法が提案されており、本論文ではこれらの方法を用いて反復計算したシミュレーションを行い、その性能の検証を行った。結果をまとめたのが表-2である。この表では、変数選択の結果のみを扱い、100回の反復計算において各独立変数が選ばれた回数をカウントしている。また、変数選択の安定性を示すために計算したJaccard係数も示している。シミュレーションの全体の結果は補足資料で示している。

表- 2 安定性選択とランダム化された安定性選択の結果

	ホールドアウト法	K分割交差検証	反復交差検証	モンテカルロ交差検証	入れ子交差検証	一つ抜き交差検証	入れ子交差検証	Adaptive Lasso
(Intercept)	100	100	100	100	100	100	100	100
displacement	97	18	100	100	100	100	100	100
foreign	100	100	100	98	100	100	100	100
gear_ratio	43	100	58	77	50	62	50	50
headroom	99	20	100	98	100	100	100	100
length	27	100	22	59	34	32	34	34
mpg	48	30	23	70	22	24	22	22
rep78	70	67	100	98	100	100	100	100
trunk	31	100	22	62	22	26	22	22
turn	55	42	100	80	74	80	74	74
weight	100	100	100	100	100	100	100	100
Jaccard係数	0.701	0.788	0.775	0.792	0.793	0.799	0.793	0.794

Jaccard係数を見ると、ホールドアウト法が0.701で最も低い。この方法は最初に考案されたものであるが、変数選択の点では不利のようだ。一つ抜き交差検証が最も高く0.799であった。ただし、Adaptive Lasso、入れ子交差検証、K分割交差検証、反復交差検証は0.77～0.80の間にあり、一つ抜き交差検証が特に高い一致度を示しているわけではない。

理論上、網羅的交差検証である一つ抜き交差検証が最も高い一致率を出すはずである。この点に関してはシミュレーションでも確認できたが、非網羅的交差検証でも遜色のない近い一致率が示された。一方で、最も一致率の高い一つ抜き交差検証でも、毎回同じ変数が選ばれるわけではなく、再現性の点では、一つ抜き交差検証でも

不足があるという結果であった。

3.2. 安定性選択の結果

計算結果を示したのが表-3である。

表- 3 安定性選択とランダム化された安定性選択の結果

	安定性選択	ランダム化された安定性選択
weight	0.78	0.75
displacement	0.77	0.77
foreign	0.92	0.92

どちらの方法でも「foreign」「weight」「displacement」の3つが選ばれている。交差検証の反復計算シミュレーションでも、これらの3つの変数は100回中100回選ばれることが多く、交差検証の計算方法との整合性も確認された。

前進選択法とモデル経路選択の結果は図-2と図-3に示す。

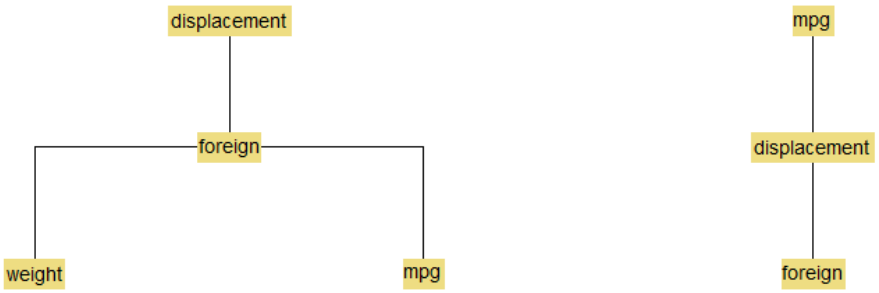


図- 2 前進選択法の結果

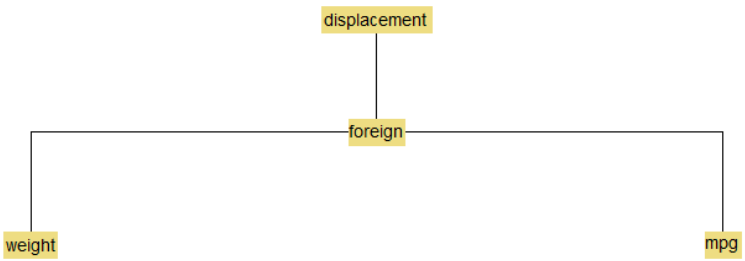


図- 3 モデル経路選択の結果

各変数の関係性も示された図であるが、Lasso 回帰では独立変数間の関係はあまり重要ではないため、「displacement」、「foreign」、「weight」、「mpg」の 4 つの変数が選ばれたという結果になる。これは、安定性選択やランダム化された安定性選択よりもスパースではない結果となっている。

安定性選択、ランダム化された安定性選択、前進選択法、モデル経路選択のどれを選択すべきかは、統計手法やシミュレーションだけでは一意に決められない。解析するデータの特性や、分析者がモデルをスパースにしたいかどうかによって、適切な方法を選択することになる。

4. 議論

4.1. 安定性選択の後に Ridge 回帰を用いて推定する

変数選択という観点から見ると、交差検証の一致度は 0.77～0.80 であり、ある程度の一致が確認できたが、統計技法として再現性が高いとは言えなかった。本論文の目的は、変数選択と推定の安定性を達成することであるため、安定性選択を用いるのが適していると考えられる。以後、「安定性選択」という用語を使用するが、本論文では安定性選択、ランダム化された安定性選択、前進選択法、モデル経路選択の 4 つの技法を代表するものとして使用することにする。

安定性選択は変数選択のみを行う。したがって、安定性選択後には選ばれた変数を用いて Lasso 回帰か Ridge 回帰を実施する。Lasso 回帰のアルゴリズムには変数選択が含まれているため、安定性選択で選ばれた変数が Lasso 回帰で除外される可能性はある。ただ、安定性選択はスパースなモデルを提示するため、多くの場合、安定性選択後に変数がさらに削減される可能性は低い。本論文のシミュレーションでも、安定性選択で選ばれた変数はすべて Lasso 回帰でも選ばれていた。

一方で、変数選択がすでに行われているとすれば、変数選択がアルゴリズムに含まれる Lasso 回帰ではなく、変数選択を行わない Ridge 回帰(Hoerl & Kennard, 1970)を用いるのも選択肢の一つになる。Ridge 回帰を選択する積極的な理由の一つは、Lasso 回帰では対処できない多重共線性を除外できることである。多重共線性の判断は一般的に VIF(Variance Inflation Factor)は使われるが、カットオフの問題(O'brien, 2007)や VIF の限界が指摘されている(Salmerón et al., 2020)。VIF などの指標が全く役に立たないわけではないが、VIF だけで結論を出すことには問題がある。したがって、VIF を参考にしつつ、独立変数が似通っている場合や Lasso 回帰と大きく異なる推定値を Ridge 回帰が出す場合には、Ridge 回帰を選択すべきである。

しかし、Ridge 回帰にも問題がないわけではない。Ridge 回帰も Lasso 回帰と同じく乱数を用いてλを計算するため、計算のたびに異なる結果になるのは Lasso 回帰と同様である。そこで、安定性選択後に Ridge 回帰を行うアルゴリズムの性能を評価するため、異なる乱数シードを設定して 100 回の反復計算を行い、推定の安定性についてシミュレーションを実施した。

本論文のシミュレーションデータを用いた Ridge 回帰では、Lasso 回帰のように選ばれるλが毎回異なることはなく、比較的安定してλが選ばれる傾向にあった。100 回の計算で選ばれたλを集計した結果を以下の表- 4 に示す。

表- 4 Ridge 回帰で選ばれたλの回数

λ	回数
158.4189421	72
173.8645934	9
190.8161766	7
209.4205182	7
229.8387602	2
252.2477556	2
333.4566878	1

100 回中 72 回は「158.4189421」がλとして選ばれており、「158.4189421」をλ値として指定して推定すると推定値は一意に決まる。選ばれるλの回数からλを選ぶという方法も選択できる方法の一つである。

λ値を指定しない場合は「158.4189421」以外のλ値も選ばれることから、100 回の反復の推定値の平均とその標準偏差を計算し、λ=158.4189421 の時の値も計算し表- 5 に示した。

表- 5 100 回の Ridge 回帰の平均値と標準偏差とλを指定した時の推定値の比較

	100 回の反復計算		λ値固定
(Intercept)	-2619.728	(123.15)	-2677.703
weight	1.689	(0.03)	1.701
displacement	13.363	(0.13)	13.425
foreign	3278.838	(69.01)	3311.322

100 回中 72 回も同じλが選ばれた結果でもあるが、両者は誤差といえる程度の推定値の違いしかないことが判明した。本論文のシミュレーションでは、推定値の安定を考えれば、どちらの方法をとってもよいと考えられる結果となった。

4.2. 安定性選択の後に Lasso 回帰を用いて推定する

4.2.1. アンサンブル法とブートストラップ法の比較

安定性選択をしているため変数の選択は一意に決定されるものの、Lasso 回帰を行う際には毎回異なる乱数を用いて計算するため、毎回異なる λ を使用することになり、その結果、推定値は毎回異なる数値が計算される。

本論文では、変数の選択の安定性と推定の安定性を達成することを目的としている。したがって、安定性選択を用いるだけでは本論文の目的を達成することはできない。ランダム性によって結果が変わる場合、統計学ではアンサンブル法やブートストラップ法を用いるのが推奨されている(Dikta & Scheer, 2021)。つまり、1 回の計算では推定値は安定しないものの、何回も計算すればランダム性による変動を制御でき、真の推計値に近い値を求められるという考え方である。

そこで本論文では、アンサンブル法とブートストラップ法(Hastie et al., 2009)の 2 つの方法を用いて推定値の平滑化のシミュレーションを行う。どちらの方法も反復回数は今までのシミュレーションと同じく 100 回行う。また、アンサンブル法では、今までのシミュレーションと同じく計算のたびに異なる乱数シードを与え、100 回同じ方法で推定し、その平均値と標準偏差を取る方法を採用した。交差検証の方法は標準的な 10-fold 交差検証を用いた。

アンサンブル法とブートストラップ法によって、100 回の反復計算をした推定値の平均と標準偏差を示したのが表- 6 ある。

表- 6 アンサンブル法・ブートストラップ法の推定の安定性のシミュレーション結果

	アンサンブル法		ブートストラップ法	
(Intercept)	-3382.08	(202.54)	-3460.23	(1915.79)
weight	1.87	(0.04)	1.92	(0.88)
displacement	13.82	(0.27)	13.33	(7.55)
foreign	3689.20	(116.39)	3695.02	(726.98)

表- 6 ではアンサンブル法やブートストラップ法のモデルの安定性を評価する方法として推定値の標準偏差で表している。標準偏差が少ないということは、各計算での推定値がほぼ同じであることを示しており、数値が小さいほどモデルの安定性が高いことを意味する。アンサンブル法とブートストラップ法を比較した結果、全ての変数においてアンサンブル法の標準偏差が小さく、シミュレーション結果からアンサンブル法が優れていることがわかる。この結果を図示したものが図- 4 と図- 5 図- 5 ある。

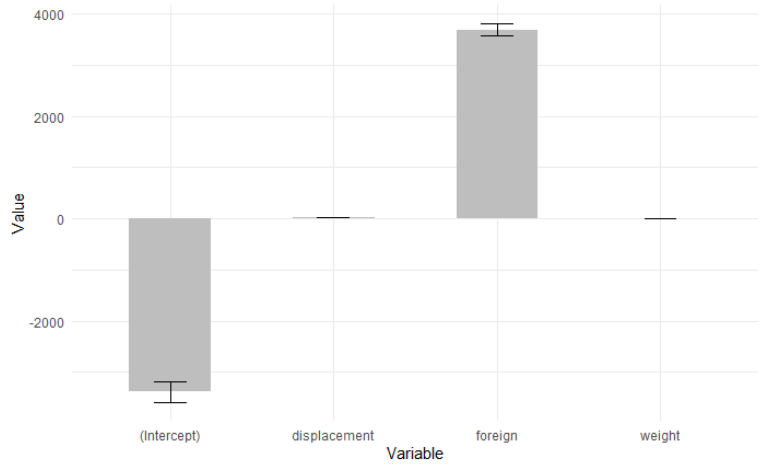


図- 4 アンサンブル法を用いて平滑化した推定値と標準偏差

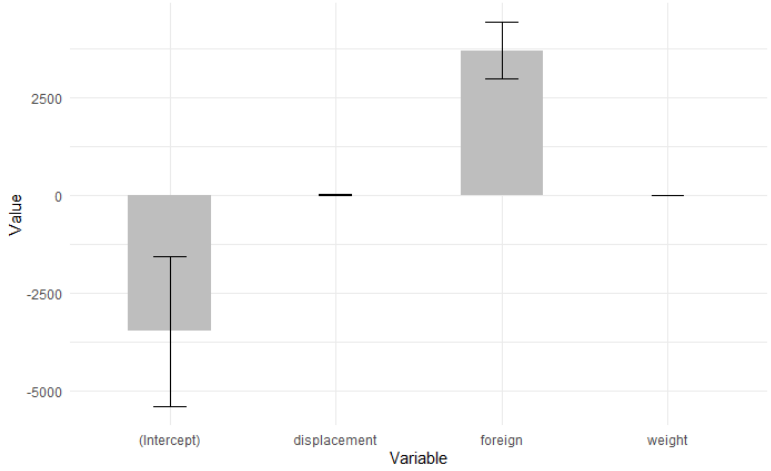


図- 5 ブートストラップ法用いて平滑化した推定値と標準偏差

この結果はブートストラップ法のリサンプリングを行うというアルゴリズムに起因すると考えられる。リサンプリングして各推定で使用するケースが異なると、推定値のバラツキが生じるため、標準誤差が大きくなったのだと考えられる。

4.2.2. 交差検証ごとの推定量の安定性のシミュレーション結果

アンサンブル法とブートストラップ法のシミュレーションでは、Lasso 回帰で最も標準的に使われている 10-fold 交差検証を用いて推定を行った。本論文で検討してきた交差検証の方法を使用することもできるため、他の方法でのシミュレーションしたのが表- 7 表- 4 である。

表- 7 アンサンブル法を用いて平滑化した各モデルの推定値の平均とその標準偏差

	10-fold 交差検証		反復交差検証		モンテカルロ交差検証	
(Intercept)	-3382.08	(202.54)	-2821.62	(184.95)	-2674.42	(350.50)
weight	1.87	(0.04)	1.74	(0.04)	1.71	(0.07)
displacement	13.82	(0.27)	13.50	(0.20)	13.34	(0.37)
foreign	3689.20	(116.39)	3389.26	(103.25)	3307.03	(195.58)
	入れ子交差検証		一つ抜き交差検証		Adaptive Lasso	
(Intercept)	-3324.01	(290.92)	-3263.43	(483.69)	-2843.68	(941.19)

weight	1.86	(0.05)	1.85	(0.09)	1.47	(0.54)
displacement	13.75	(0.39)	13.67	(0.63)	17.20	(3.73)
foreign	3656.14	(166.73)	3621.99	(276.43)	3719.79	(121.72)

先ほど行った 10-fold 交差検証も標準偏差が小さいモデルとなっているが、わずかに反復交差検証のモデルの標準偏差が小さくなっていることが確認できる。一方で、モンテカルロ交差検証と Adaptive Lasso は標準偏差が大きく、あまり適していない可能性が示唆された。

なお、標準偏差が少なく精度の高い推定ができることと、データを精度高く推定できることは、また別の問題である。そこで表-7の推定量の平均値を基にモデルの性能を MSE、MAE、RMSE、MAPE を計算したのが表-8である。

表- 8 アンサンブル法を用いて平滑化した各モデル性能評価指標の値

	MSE	MAE	RMSE	MAPE
10-fold 交差検証	**4006434**	**1508.786**	**2001.608**	25.39876
反復交差検証	4029164	1511.283	2007.278	25.22268
モンテカルロ交差検証	4051043	1507.493	2012.72	**24.98080**
Adaptive Lasso	4019477	1512.033	2004.863	25.56368
一つ抜き交差検証	4009386	1510.757	2002.345	25.39737
入れ子交差検証	4007704	1509.754	2001.925	25.39826

MAPE を除いて、他の指標では 10-fold 交差検証が最も良いモデルとなった。モンテカルロ交差検証および Adaptive Lasso 以外の方法の結果を見ると、大きなモデル性能の差異は見られない。したがって、モンテカルロ交差検証と Adaptive Lasso を除けば、どの方法を選んでも大きな違いは生じない結果となった。ただし、K 分割交差検証はモデル性能が良好であり、かつ計算コストが最も低いため、最も有力な選択肢となるであろう。

4.2.3. アンサンブル法の必要性の評価

推定値の平滑化に関して、本論文のシミュレーション結果ではアンサンブル法が有効であることが示された。最後に実施するシミュレーションでは、アンサンブル法を用いることで推定値がどの程度平滑化されるか、またその必要性を検証する。この結果、アンサンブル法が推定値の平滑化に大きく貢献するのであれば、その積極的な活用が推奨される。しかし、アンサンブル法を用いない場合と大きな違いがなければ、計算コストの高いアンサンブル法を採用する必要はないと考えられる。各交差検証で1回だけ計算した結果を表-9以下の表に示す。

表- 9 安定性選択後に各交差検証を用いた Lasso の結果

	10-fold 交差検証	反復 交差検証	モンテカルロ 交差検証	入れ子交差 検証	一つ抜き 交差検証	Adaptive Lasso
(Intercept)	-3369.49	-2868.12	-2093.23	-3520.06	-3369.49	-3429.25
weight	1.87	1.75	1.59	1.89	1.87	1.81
displacement	13.79	13.55	12.73	14.02	13.79	14.80
foreign	3680.98	3415.21	2982.63	3769.20	3680.98	3790.59

表-7と表-9を見比べるとどの程度の違いがあるかある程度わかるが、数値を逐一見ることは不正確であるため、100回の反復計算を行った結果をスタンダードとして1回だけの計算で、どの程度離れているかを全体Zスコアと平均相対誤差を用いて計算した結果を表-10に示す。

表– 10 アンサンブル法と 1 回の計算を比較した統計量

	全体 Z スコア	平均相対誤差
10-fold 交差検証	0.244	0.002
反復交差検証	0.884	0.008
モンテカルロ交差検証	6.734	0.108
入れ子交差検証	2.348	0.032
一つ抜き交差検証	0.825	0.017
Adaptive Lasso	2.479	0.149

表- 10 では 10-fold 交差検証が最も良いという結果が示されている。表- 11 はアンサンブル法と 1 回の計算結果の比較である。概ね結果は一致しているものの、4 桁の「切片」は 2 桁目に誤差が現れており、2 桁の「displacement」は小数点以下第 1 位に誤差が現れている。この誤差が問題となる場合には、アンサンブル法の使用を検討するのが望ましいと考えられる。

表– 11 10–fold 交差検証のアンサンブル法と単回推定の比較

	アンサンブル法		単推定
(Intercept)	-3382.08	(202.54)	-3369.49
weight	1.87	(0.04)	1.87
displacement	13.82	(0.27)	13.79
foreign	3689.20	(116.39)	3680.98

5. 結論

本論文では、交差検証のシミュレーションを行った下記の知見が得られた。

A) 変数の安定的な選択と推定という点では、交差検証の方法を変えても大きな違いはなかった。Jaccard 係数で 0.77～0.80 の一致度であり、安定しているとは言い難い結果であった。Adaptive Lasso に関しても大差はなかった。
B) 安定性選択を分析に組み込むことで安定した変数選択が可能になった。

この結果を踏まえ、本論文では、安定性選択を推奨する。
次に問題になるのが安定性選択後のアルゴリズムをどのようにするかということである。一つは Ridge 回帰を行うという方法である。Ridge 回帰は多重共線性を抑制できる利点があり、特に独立変数が類似している場合には積極的に Ridge 回帰を選ぶべきである。Ridge 回帰では 2 つのアルゴリズムを提案する。

1) 異なる乱数を与えて Ridge 回帰を 100 回反復計算し、最も多く選ばれたλを選択する。この方法の利点は推定値が一意に決まるところにある。
2) 異なる乱数を与えて Ridge 回帰を 100 回反復計算し推定値の平均を求める。

本論文のシミュレーションの結果からは、1 と 2 の方法のどちらをとっても良いと考えられる。
次に Lasso 回帰のアルゴリズムを提案した。シミュレーションでは、アンサンブル法とブートストラップ法を比較するとブートストラップ法は結果の標準偏差が大きく本論文のテーマには適していないことが分かった。

3) 6 つの交差検証とアンサンブル法用いた Lasso 回帰を行った結果、モンテカルロ交差検証と Adaptive Lasso は適していないことがわかった。K 分割交差検証はモデル性能が良好で、かつ計算コストが最も低いため、本論文は K 分割交差検証とアンサンブル法の組み合わせを推奨する。

4) 推定値の安定のためにアンサンブル法を用いたが、単回推定と大差がなく、必ずしもアンサンブル法は必須ではない。安定性選択後にK分割交差検証のLasso回帰を行うことも多くの場合許容範囲にあり、どの程度の誤差が許容できるかを判断して使用することが必要である。

本論文では、変数の推定値の安定性という観点から、以上の4つのアルゴリズムを提案する。

本論文の限界は、実データを用いたシミュレーションを行うという方法論にある。Lasso回帰にはアルゴリズムに乱数が使用されているため、Lasso回帰の安定性を理論的に評価することが困難であるため、データを用いたシミュレーションを行った。しかし、この方法にはシミュレーションに使用したデータに大きく依存するという欠点がある。つまり、本論で使用したデータの特性がシミュレーション結果に影響を与えている可能性がある。この点を克服するためには、様々なデータを用いてシミュレーションを繰り返すことが唯一の解決策となる。

Lasso回帰は非常に強力なツールである一方で、この再現性の問題において困難性を抱えている。乱数を使用するLasso回帰では、変数選択と推定値が計算のたびに異なるという問題が存在する。科学において再現性は重要であり、統計解析においても再現性は非常に重要である。この問題を解決するために、本論文では実データを用いたシミュレーションを実施し、再現性を実現するアルゴリズムを提案した。

本論文は統計解析のシミュレーションにとどまっているが、提案したアルゴリズムを使用すれば、機械学習、社会科学、生物学、医学など多くの分野で応用が可能であり、再現性を確保した研究を促進することができるであろう。これにより、研究の信頼性と正確性が向上し、実際の応用においても重要な役割を果たすことが期待される。さらに、これらの分野におけるデータ解析の効率化と精度向上にも寄与するため、学術界や実務において広範な貢献をすることができるだろう。

補足資料

本論文で使用したRのコードなど補足資料は下記からダウンロードできる。
https://zenodo.org/doi/10.5281/zenodo.12594772

引用文献

Afreixo, V., Tavares, A. H., Enes, V., Pinheiro, M., Rodrigues, L., & Moura, G. (2024). Stable Variable Selection Method with Shrinkage Regression Applied to the Selection of Genetic Variants Associated with Alzheimer's Disease. *Applied Sciences*, *14*(6), 2572. https://doi.org/10.3390/app14062572

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(none). https://doi.org/10.1214/09-SS054

Bishop, C. M. (2016). *Pattern Recognition and Machine Learning* (Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009)). Springer New York.

Dikta, G., & Scheer, M. (2021). *Bootstrap Methods: With Applications in R*. Springer International Publishing. https://doi.org/10.1007/978-3-030-73480-0

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, *70*(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

He, Q., Zhang, M., Zhang, J., Yang, S., & Wang, C. (2023). K-Fold Cross-Valuation for Machine Learning Using Shapley Value. In L. Iliadis, A. Papaleonidas, P. Angelov, & C. Jayne (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2023* (Vol. 14256, pp. 50–61). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44213-1_5

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin*

de La Société Vaudoise Des Sciences Naturelles, *37*(142), 547. https://doi.org/10.5169/seals-266450

Kissel, N., & Mentch, L. (2024). Forward stability and model path selection. *Statistics and Computing*, *34*(2), 82. https://doi.org/10.1007/s11222-024-10395-8

Kohavi, R. (1995, August 20). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. International Joint Conference on Artificial Intelligence. https://www.semanticscholar.org/paper/A-Study-of-Cross-Validation-and-Bootstrap-for-and-Kohavi/8c70a0a39a686bf80b76cb1b77f9eef156f6432d

McLachlan, G. J., Do, K., & Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data* (1st ed.). Wiley. https://doi.org/10.1002/047172842X

Meinshausen, N., & Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *72*(4), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307. https://doi.org/10.1093/bioinformatics/bti499

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Salmerón, R., García, C., & García, J. (2020). *Overcoming the inconsistences of the variance inflation factor: A redefined VIF and a test to detect statistical troubling multicollinearity* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2005.02245

Shah, R. D., & Samworth, R. J. (2013). Variable Selection with Error Control: Another Look at Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *75*(1), 55–80. https://doi.org/10.1111/j.1467-9868.2011.01034.x

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(1), 44–47. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Xu, Q.-S., & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

馬場真哉. (2018). *Python で学ぶあたらしい統計学の教科書*. 翔泳社.

(2024年8月28日受理)