データ解析と統計学

廣瀬　慧 2019.8.11

1　はじめに

近年，機械学習や統計解析のソフトウェアが普及し，データ解析が身近なものになりました．予測精度の高い電力需要予測は，エネルギー経済の活性化・再エネ促進へとつながります．遺伝子データ解析は，病気の予測に役立ちます．地面の下の動きを分析することにより，地盤の硬さを推定でき，土砂災害の予測へと繋がります．データ解析は人の役に立つ有用なツールとして，今後ますます必要になると考えられます．

世の中に統計解析手法は数多く存在しますが，あらゆる統計解析手法の基礎となっているのが，線形回帰モデルです．このテキストでは，線形回帰モデルとモデルに含まれるパラメータの推定法について解説します．

2　線形回帰モデル

回帰モデルは，p 次元説明変数ベクトル $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ から目的変数 y を予測するためのモデルです．ここで，“A^T” は行列 A の転置を表します．なお，$\boldsymbol{x}$ を p 次元と書きましたが，y を予測するための変数の数が p 個あるとイメージしていただけると分かりやすいかと思います．たとえば，遺伝子データ解析では，$\boldsymbol{x}$ の各成分は遺伝子，y は病気の進行度（たとえばがんの大きさ）などを表します．回帰モデルを構築することにより，遺伝子データから，病気がどのくらい進行するのか，予測できるようになります．

線形回帰モデルは，

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

で表されます．ここで，$\beta_0, \beta_1, \ldots, \beta_p$ はパラメータ，ε は誤差で，分布 $\varepsilon \sim (0, \sigma^2)$ に従うとします．σ^2 は分散パラメータで，誤差のばらつきの大きさを表します．

我々は，回帰モデルのパラメータ $\beta_0, \beta_1, \ldots, \beta_p$ の値を知りません．そのため，これらのパラメータをデータから推定する必要があります．観測されたデータ $\{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}$ から推定する方法として，最小二乗法と呼ばれる方法があります．最小二乗法では，誤差の二乗和

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\}^2$$

の最小化によってパラメータ $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_p)$ を求めます．ここで，$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ です．実際，行列とベクトルを用いることにより，$\boldsymbol{\beta}$ の最小二乗推定値 $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ は

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (X^T X)^{-1} X^T \boldsymbol{y} \tag{1}$$

と陽に求まります*．ただし，

$$X = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n)^T,\ \boldsymbol{y} = (y_1, \ldots, y_n)^T$$

とします．ここで，$\tilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^T)^T$ とします（“1”は切片項に対応します）．

ここで，新たな説明変数ベクトル $\boldsymbol{x}$ が得られたとき，さきほど推定したパラメータ $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ を用いて

$$\hat{y} = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$$

として y の値を「予測」できます．この予測は，どんな p 次元説明変数ベクトル $\boldsymbol{x}$ に対しても適用できますので，$\boldsymbol{x}$ さえ与えられればすぐに予測ができるようになるのです．

線形回帰モデルは「万能なツール」に見えます．しかしながら，実際はそうではありません．なぜなら，線形回帰モデルは,「y が $x_1, \ldots, x_p$ の線形結合で近似できる」という仮定を置いているからです．この仮定を満たさなければうまくいきません．統計では，何らかの「仮定」をおいて，その仮定が満たされたときの理論構築を行うことが多いのですが，実際問題，その仮定が現実で満たされるのかどうかわからないことが多いです．

しかし,「この仮定は正しそう」ということはわかります．たとえば，データをプロットして線形の関係性にあれば，なんとなく正しいと言えそうです．さらに，仮説検定を使うと,「統計的にこの仮説は有意に正しくない」というように，ある仮説が正しいかどうかを統計的に議論することができます．もちろん,「100%この仮定は成り立っている」ということはいえませんが，自信を持って「正しそうである」といえるのは重要なことだと思います．

また，仮に線形であるという仮定が満たされたとしても，一般に，サンプルサイズが小さく，不要な説明変数の数が大きくなると，予測誤差が大きくなるという問題が生じます．

3 Lasso

近年，説明変数の個数 p は増えてきており，たとえば，遺伝子データだと，p が数万になることがあります．このとき，最小二乗法を求めることが難しくなります．たとえば，サンプルサイズが説明変数の個数より小さいとき，すなわち $n < p$ のときは，$X^T X$ は逆

*厳密には，$X^T X$ が逆行列を持つという条件が必要です．

行列を持たず，最小二乗推定値を計算できません．また，データが大規模になれば，そもそもパラメータを計算機で計算するということ自体が困難になることも多々あります．

これらの問題を対処する方法として，正則化法というのがあります．とくに，Tibshirani (1996, JRSSB) の提案した Lasso (Least absolute shrinkage and selection operator) は，p が大きい場合でも用いることができる方法として広く用いられています．Lasso 推定値は以下の関数を最小化することにより求められます．

$$\sum_{i=1}^{n}\left\{y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)\right\}^2+\lambda\sum_{j=1}^{p}|\beta_j|.$$

ただし，$\lambda\geq 0$ は正則化パラメータとします．最小二乗法と異なる点は，上の式の第 2 項目にあります．じつは，λ の値が大きければ，$\boldsymbol{\beta}$ の成分のち，いくつかが 0 に推定されることが知られています．0 であるということは，対応する変数は目的変数に影響を与えないということを意味します．つまり，lasso を使うと，不要な変数を自動的に除去し，必要な変数のみピックアップできるのです．

Lasso の目的関数には，パラメータの絶対値に基づく項 $|\beta_j|$ が存在し，この項は $\beta_j=0$ で微分不可能です．そのため，数学的に扱いにくいです．たとえば，(1) 式にあるように推定値を解析的に求めるということが困難となります．しかしながら，2000 年あたりから計算アルゴリズムと高次元の理論が発展し，微分不可能でも実際に PC で動かすと有用であることがわかってきました．ここでは計算アルゴリズムと高次元の理論について少し触れたいと思います．

3.1　計算アルゴリズム

いま，λ を与えたとき，$\mathcal{A}=\{j|\hat{\beta}_j\neq 0\}$ と定義します．このとき，Lasso の目的関数を微分して[†]解くと，$\hat{\boldsymbol{\beta}}$ は次で与えられます．

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}=(X_{\mathcal{A}}^TX_{\mathcal{A}})^{-1}\left\{X_{\mathcal{A}}^T\boldsymbol{y}-\frac{\lambda}{2}\boldsymbol{s}_{\mathcal{A}}\right\} \tag{2}$$

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}^C}=\mathbf{0} \tag{3}$$

ただし，$\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ は，推定値 $\hat{\boldsymbol{\beta}}$ のうち $\mathcal{A}$ に対応する要素のみを取り出したベクトルです．

ここで，λ を大きい値から徐々に小さくしていったときの推定値 $\hat{\boldsymbol{\beta}}$ がどのように変化していくのかを見てみます．(2) 式より，Lasso 解は区分線形であることがわかりますが，すべての λ に対する $\hat{\boldsymbol{\beta}}$ をもとめるために，もうすこし詳しく調べてみます．まず，十分大きな λ に対して，$\hat{\boldsymbol{\beta}}=\mathbf{0}$ です．$\hat{\boldsymbol{\beta}}=\mathbf{0}$ となる最小の λ を λ_0 とすると，(2) 式より，

[†]原点で微分不可能なので，正確には劣微分という概念を使います．

$\lambda_0 = \max_j |\boldsymbol{x}_j^{*T}\boldsymbol{y}|/\boldsymbol{x}_j^{*T}\boldsymbol{x}_j^*$ で与えられます．ただし，$\boldsymbol{x}_j^*$ は X の j 列目を取り出した n 次元ベクトルとします．$j_1 = \arg\max_j |\boldsymbol{x}_j^{*T}\boldsymbol{y}|/\boldsymbol{x}_j^{*T}\boldsymbol{x}_j^*$ とすると，λ を λ_0 より少しだけ小さいときの推定値は，$\mathcal{A} = \{j_1\}$ としたときの (2) 式で与えられます．さらに λ を小さくすると，ある $j_2 \neq j_1$ に対し，あらたな非ゼロ推定値 $\hat{\beta}_{j_2}$ が出現する λ_1 が存在します．このとき，$\mathcal{A} = \{j_1, j_2\}$ と更新します．λ が λ_1 より少しだけ小さいときの推定値は，$\mathcal{A} = \{j_1, j_2\}$ としたときの (2) 式で与えられます．

以下，同様にして変数を追加または削除していけば良いということになります．変数が追加または削除される瞬間の λ を λ^* とすると，じつは，λ^* は陽に求めることができます．また，先に述べたように，λ を与えたとき，推定値は (2) 式，(3) 式で与えられます．つまり，任意の λ に対する推定値が得られるということになります．

3.2　収束レート

いま，パラメータのほとんどが 0 であるとします．パラメータ $\boldsymbol{\beta}$ のうち，非ゼロ要素に対応する添え字を S_0 とし，$\boldsymbol{\beta} = (\boldsymbol{\beta}_{S_0}^T, \boldsymbol{\beta}_{S_0^C}^T)^T$ とします．また，S_0 の大きさ（非ゼロ要素の数）を s_0 とします．

いま，予測精度と同等の次の二乗誤差に基づくリスク

$$R = \frac{1}{n}\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|_2^2$$

を評価することを考えます．この値が，n が大きくなるにつれて 0 に近づけば，予測値が（平均的に）正しい値に収束します．したがって，n が十分大きいとき，R が 0 に近づくことを示せば，精度よく予測できると言えます‡．最小二乗推定量に対する二乗誤差リスク R の期待値は，

$$E[R] = \frac{1}{n}E\|X(X^TX)^{-1}X^T\boldsymbol{\varepsilon}\|_2^2 = \frac{p}{n}\sigma^2 \tag{4}$$

で与えられます§．これより，p が大きいときは，$E[R]$ は大きくなってしまいます．たとえば，$p = n/2$ とすると，$E[R] = \sigma^2/2$ となり，$n \to \infty$ としても 0 に収束しません．

仮に真の非ゼロ集合 S_0 を知っているとして，最小二乗法によって推定したとすると，

$$E[R] = \frac{1}{n}E\|X\hat{\boldsymbol{\beta}}_{S_0} - X\boldsymbol{\beta}\|_2^2 = \frac{s_0}{n}\sigma^2 \tag{5}$$

‡誤差 ε が確率変数であるため，R も確率変数となります．そのため，厳密には，$\lim_{n\to\infty} R = 0$ を示すということはできません．正確には，確率収束や概収束という，普通の数列とは違った収束を考える必要があります．

§厳密には $n < p$ のときに最小二乗推定値は存在しないため，ここでは $n > p$ で p が n に十分近いときを考えています．

となります．よって，$s_0 \ll p$ のとき（すなわち十分にスパースなとき）リスクの期待値 $E[R]$ は (4) 式と比べて十分に小さくなります．もちろん，我々は集合 S_0 を知らないので，$\hat{\boldsymbol{\beta}}_{S_0}$ を求めることは不可能であり，(5) 式は,「理想的な状況下で得られた推定量に対するリスク」と解釈できます．また，(5) 式の右辺は，$n \to \infty$ としたときに 0 に収束しますので，

$$\lim_{n\to\infty} E[R] = 0 \tag{6}$$

が成り立ちます．そこで，(5) 式のリスクに近づけることを目標とします．

最小二乗法は (6) 式が成り立ちませんが，じつは，Lasso はそれに近い結果が得られます．ただ，計画行列 X に対して制約条件を課す必要があります．代表的な条件が，compatibility condition と呼ばれる条件で，$X^T X$ の最小固有値が小さすぎないという条件とみなすことができます¶．このとき,（かなりラフに書くと），n が十分大きいとき，高い確率で

$$R \leq C\frac{\log p}{n} \tag{7}$$

が成り立ちます．ここで，C は n, p に依存しない定数です．

(7) 式を見ると，右辺が (5) 式よりは大きい値をとっているものの，(4) 式の値よりは明らかに小さいです．実際，p が十分に大きいとき，分子の p と $\log p$ の差は大きいです．たとえば，$p = 10000$ のとき，$\log p \approx 9.21$ です．たとえば，p を n に依存させて，$p_n = n^k$ $(k > 1)$ とすると，(7) 式の右辺は 0 に収束します．これより，(6) 式に近い性質が成り立つのです．最小二乗法（すなわち (4) 式）ではこの性質は全く成り立ちません．

4　おわりに

Lasso は効率的なアルゴリズムが存在し，かつ $n < p$ のときに統計的に良い性質を持ちます．それゆえ，Lasso は有用なツールであり，実際，様々な場面で使われます．しかしながら，Lasso はいつも $n < p$ のときにうまくいくとはかぎりません．というのも，Lasso がうまく機能するためには，(7) 式の compatibility condition を満たす必要があるからです．じつは，この条件はとても強く，現実の場面ではこの条件を満たしている状況はあまりないと言ってもいいかもしれません．たとえば，遺伝子データに対しては，遺伝子間には大きな相関があり，Lasso はうまく機能しない可能性が高いです．一方で，画像データの復元は，変数間に大きな相関関係があまりなく，Lasso がうまく機能することが多いと考えられます．実際，Lasso は画像データ解析との相性が良いです．このように，手法の特徴を理解しておくことが，よりよい解析へとつながります．

¶実際にはもう少しゆるい条件です．

A 文献

統計解析や回帰分析の基礎については，多くの数理統計学の本に書かれており，例えば以下の本には，回帰分析について比較的詳しく書かれています．

[1] 稲垣宣生 (2003)．数理統計学．裳華房．

Lasso のアルゴリズムや高次元の理論は以下の本に詳しく書かれています．

[2] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall/CRC.

また，プログラム R を使いながら様々な解析手法について説明した，より実践的な本として，

[3] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York: Springer.

があります．