スパースな状況 $(p > N)$ における問題

- 線形回帰で最小二乗法の解が求まらない；与えられた行列 $X \in \mathbb{R}^{N \times p}$ に対し，解を求めるのに必要な $X^t X$ の逆行列が存在しない.

- 情報量基準を用いて複数の変数を探すのが困難；変数が多すぎると，推定値がランダムな挙動に対して過敏になりすぎてしまう．また，推定値を定める変数を選ぶにも p 個の変数からの選び方 $(= 2^p$通り$)$ を比較する必要がある.

これらを解決するための方策

- 二乗誤差の最小ではなく，それに係数の値が大きくなりすぎないようにするための正規化項を加えた関数の最小化問題を考える．これによって変数のノイズによって値が大きく変動してしまうことを防ぐ

- 正規化項としては係数の $L1$ ノルムの定数 λ 倍である Lasso, 係数の $L2$ ノルムの定数 λ 倍である Ridge とよぶ.
  (どうしてこれらについて考えるか$\rightarrow$ Lasso ,Ridge は共に凸関数の和であることから凸関数となり，最小化問題について考えやすくなる)

- Lasso がモデル選択の役割を持つ$\rightarrow$ Lasso では λ を大きくすることで特定の係数を 0 にすることができ，それによって注目すべき説明変数を絞ることができる

考えている状況

N 個のデータ

$$\begin{gathered}(x_{1,1}, \cdots, x_{1,p}, y_1)\\(x_{2,1}, \cdots, x_{2,p}, y_2)\\\vdots\\(x_{N,1}, \cdots, x_{N,p}, y_N)\end{gathered}$$

が与えられており，これらを基に y の値を $\boldsymbol{x} \in \mathbb{R}^p$ で推定したい.
$\Rightarrow$ 各 j に対して $\beta \in \mathbb{R}^p$,$\beta_0 \in \mathbb{R}$ を用いて

$$\hat{y_j} := \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_p x_{p,j} = \beta_0 + \langle \beta, \boldsymbol{x} \rangle$$

によって y の値を推定する．このとき，実測値 y と推定値 $\hat{y}$ との差

$$
\begin{aligned}
y-\hat{y} &= \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{pmatrix} - \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,p} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \\
&= y - \boldsymbol{\beta}_0 - X\beta
\end{aligned}
$$

の L_2 ノルム[*1]の 2 乗を最小にする切片 β_0 と傾き β を求める．

まず各 j に対して，X の第 j 列と y が中心化されているとする；
$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,p} \end{pmatrix}$ と $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$ に対して

$$
\begin{aligned}
X' &= \begin{pmatrix} x_{1,1}-\bar{x_1} & \cdots & x_{1,p}-\bar{x_p} \\ \vdots & \ddots & \vdots \\ x_{N,1}-\bar{x_1} & \cdots & x_{N,p}-\bar{x_p} \end{pmatrix} \quad \left(\bar{x_j} = \sum_{i=1}^{N} x_{i,j} \right) \\
y' &= \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{pmatrix} \quad \left(\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \right)
\end{aligned}
$$

と変形されており，

$$
\begin{aligned}
\bar{x'}_j &:= \frac{1}{N} \sum_{i=1}^{N} (x_{i,j} - \bar{x}_j) = 0 \quad \forall j \\
\bar{y}' &:= \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}) = 0
\end{aligned}
$$

となっているとする．このとき，X' と y' に関する最小二乗法の解 $(\hat{\beta}_0, \hat{\beta})$ のうち，$\hat{\beta}_0$ は 0 であることがわかる．

Proof. $\|y' - \boldsymbol{\beta}_0 - X'\beta\|^2 = \sum_{i=1}^{N} \left(y'_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2$ であり，この関数は β_0^2 の係数が 1 の β_0 の 二次式であるので

$$
\|y' - \boldsymbol{\beta}_0 - X'\beta\|^2 \text{が} \hat{\beta}_0 \text{において最小} \Rightarrow \frac{\partial}{\partial \beta} \|y' - \hat{\beta}_0 - X'\beta\| = 0
$$

が成立する．したがって

[*1] L_2 ノルムにすることで関数を滑らかにし，微分ができるようにしておく

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\beta}\sum_{i=1}^{N}\left(y'_i-\beta_0-\sum_{j=1}^{p}x'_{i,j}\beta_j\right)^2 = \sum_{i=1}^{N}\frac{\partial}{\partial\beta}\left(y'_i-\beta_0-\sum_{j=1}^{p}x'_{i,j}\beta_j\right)^2 \\
&= \sum_{i=1}^{N}2\left(y'_i-\beta_0-\sum_{i=1}^{N}x'_{i,j}\beta_j\right)\cdot(-1) = -2\left(\sum_{i=1}^{N}y_i-N\beta_0-\sum_{i=1}^{N}\sum_{j=1}^{p}x'_{i,j}\beta_j\right) \\
&= -2N\left(\bar{y'}-\beta_0-\sum_{j=1}^{p}\frac{1}{N}\sum_{i=1}^{N}x'_{i,j}\beta_j\right) = -2N\left(\bar{y'}-\beta_0-\sum_{j=1}^{p}\bar{x'}_j\beta_j\right) \\
&= 2N\beta_0 \qquad \therefore \beta_0 = 0
\end{aligned}
$$

*2 □

以下では X, y が中心化されているとする．このとき，切片 $\beta_0=0$ より $\|y-\boldsymbol{\beta}_0-X\beta\|^2=\sum_{i=1}^{N}(y_i-\sum_{j=1}^{p}x_{i,j}\beta_j)^2$ となる．この式は各 j に対して β_j^2 の係数が 1 の β_j の二次式なので

$$
\|y-X\beta\|^2\text{が}\hat{\beta}\text{において最小} \Rightarrow \text{各}\beta_j\text{に対して}\frac{\partial}{\partial\beta_j}\|y-X\hat{\beta}\|^2=0
$$

が成立する．また各 j に対して

$$
\begin{aligned}
\frac{\partial}{\partial\beta_j}\|y-X\beta\|^2 &= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{i,j}\beta_j\right)^2 \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial\beta_j}\left(y_i-\sum_{j=1}^{p}x_{i,j}\beta_j\right)^2 \\
&= \sum_{i=1}^{N}2\left(y_i-\sum_{k=1}^{p}x_{i,k}\beta_k\right)\cdot(-x_{i,j}) \\
&= -2\sum_{i=1}^{N}x_{i,j}\left(y_i-\sum_{k=1}^{p}x_{i,k}\beta_k\right)
\end{aligned}
$$

であることから

*2 中心化された X', y' に関する最小二乗法の解 $(\hat{\beta}_0, \hat{\beta})$ が元の X, y に関する解になるかは要証明.

$$
\begin{aligned}
(\mathbf{0} =) \begin{pmatrix} \frac{\partial}{\partial \beta_1} \|y - X\hat{\beta}\|^2 \\ \vdots \\ \frac{\partial}{\partial \beta_p} \|y - X\hat{\beta}\|^2 \end{pmatrix} &= -2 \begin{pmatrix} \sum_{i=1}^N x_{i,1} \left(y_i - \sum_{k=1}^p x_{i,k}\beta_k\right) \\ \vdots \\ \sum_{i=1}^N x_{i,p} \left(y_i - \sum_{k=1}^p x_{i,k}\beta_k\right) \end{pmatrix} \\
&= -2 \begin{pmatrix} x_{1,1} & \cdots & x_{N,1} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,p} \end{pmatrix} \begin{pmatrix} y_1 - \sum_{k=1}^p x_{1,k}\beta_k \\ \vdots \\ y_N - \sum_{k=1}^p x_{N,k}\beta_k \end{pmatrix} \\
&= -2X^t(y - X\hat{\beta})
\end{aligned}
$$

が得られる．したがって正方行列 X^tX が正則であるなら

$$
\begin{aligned}
X^t(y - X\hat{\beta}) = 0 &\Leftrightarrow X^tX\hat{\beta} = X^ty \\
&\Leftrightarrow \hat{\beta} = (X^tX)^{-1}X^ty
\end{aligned}
$$

となり，最小二乗法の解が求まる．特に $p = 1$ のとき，$X = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$ とすると $(X \neq O)$

$$
\hat{\beta} = \frac{\sum_{i=}^N x_i y_i}{\sum_{i=1}^N x_i^2} = \frac{\sum_{i=}^N x_i y_i / N}{\sum_{i=1}^N x_i^2 / N}
$$

となることがわかる．[*3]

以上より，中心化された X, y に対しては最小二乗法の解 $(\hat{\beta}_0, \hat{\beta})$ が $(0, (X^tX)^{-1}X^ty)$ となることがわかった．中心化されていない X, y に対しても，式変形によって中心化されたものに帰着することができる．

Proof.

$\bar{X} = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x_1} & \cdots & \bar{x}_p \end{pmatrix} \in \mathbb{R}^{N \times p}, \bar{\boldsymbol{y}} = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} \in \mathbb{R}^N$ とすると

$$
\begin{aligned}
\|y - \boldsymbol{\beta}_0 - X\beta\|^2 &= \|y - \bar{\boldsymbol{y}} + \bar{\boldsymbol{y}} - \boldsymbol{\beta}_0 - (X - \bar{X})\beta - \bar{X}\beta\|^2 \\
&= \|y' - (\bar{X}\beta - \bar{\boldsymbol{y}} + \boldsymbol{\beta}_0) - X'\beta\|^2
\end{aligned}
$$

[*3] つまり変数が 1 つなら，最小二乗法の解 β は $\frac{X \text{ と } y \text{ の共分散}}{X \text{ の分散}}$ で与えられる．

であり，最右辺は中心化された X', y' に関する最小二乗法であるので

$$
\begin{aligned}
\text{最右辺が最小} &\Rightarrow \bar{X}\beta - \bar{\boldsymbol{y}} + \boldsymbol{\beta}_0 = 0 \text{ かつ} \beta = (X'^t X')^{-1} X'^t y' \\
&\Leftrightarrow \beta_0 = \bar{y} - \sum_{i=1}^{p} x_i \beta_i \text{かつ} \beta = (X'^t X')^{-1} X'^t y'
\end{aligned}
$$

であることから X, y に関する最小二乗法の解が $(\bar{y} - \sum_{i=1}^{p} x_i\beta_i, (X'^t X')^{-1} X'^t y')$ となることがわかる． □

上では $X^T X$ が正則であることを仮定したが，一般に $N \times p$ 行列 X に対して，$N < p$ であれば p 次正方行列 $X^T X$ は正則でない

Proof. 行列 X が定める線形写像を $f : \mathbb{R}^p \to \mathbb{R}^N$，$X$ の転置行列 X^T が定める線形写像を $f' : \mathbb{R}^N \to \mathbb{R}^p$ とおく．このとき，$X^T X$ が正則となるためには合成写像 $f' \circ f$ に対して $\mathrm{rank}(f' \circ f) = p$ となる必要がある．
今，包含関係

$$
\mathrm{Im} f(\mathbb{R}^p) \subset \mathbb{R}^N
$$

より

$$
\mathrm{Im}(f'(f(\mathbb{R}^N))) = \mathrm{Im}(f' \circ f) \subset \mathrm{Im} f'
$$

が成立することから

$$
\mathrm{rank}(f' \circ f) \leq \mathrm{rank} f'
$$

がわかり，これと転置行列のランクが元の行列のランクに等しいことから

$$
\mathrm{rank}(f' \circ f) \leq \mathrm{rank} f \tag{1}
$$

が得られる．さらに写像 f について，次元定理より

$$
\begin{aligned}
p = \dim \mathbb{R}^p &= \mathrm{rank} f + \mathrm{null} f \\
\therefore \mathrm{rank} f &= p - \mathrm{null} f \leq p
\end{aligned} \tag{2}
$$

が成立し，また包含関係 $\mathrm{Im} f(\mathbb{R}^p) \subset \mathbb{R}^N$ より

$$
\mathrm{rank} f \leq \dim \mathbb{R}^N = N \tag{3}
$$

がわかる．したがって (2),(3) 式より

$$
\mathrm{rank} f \leq \min\{N, p\} \leq N < p
$$

が成立するのでこれと (1) 式より

$$
\mathrm{rank}(f' \circ f) \leq \mathrm{rank} f < p
$$

となるので $\mathrm{rank}(f' \circ f) < p$ がわかる．したがって行列 $X^T X$ は正則でないことが示せる．

□

また，X に同じ列が 2 個ある場合にも $(\exists i, j \in \{1, \cdots, p\}\ s.t.; i \neq j \wedge x_{k,i} = x_{k,j}\ (\forall k \in \{1, \cdots, N\}))$，$\mathrm{rank}X < p$ となるので $\mathrm{rank}X^T X < p$ となり，逆行列が存在しないことがわかる．
さらに変数の個数 p の値が大きいと，目的変数 y を説明するための説明変数を選ぶ際に 2^p 通りの変数の組み合わせを考える必要があり，計算量が膨大になる．

クロスバリデーション：いくつかのデータを分割して推定し，適切なモデルを選択する
(上の線形回帰での例)

(1) 2^p 通りの説明変数の組み合わせそれぞれに番号を振り，s 番目の組み合わせを選んだとする．(その時の変数は t 個であったとする)

(2) 与えられた N 個のデータを k 個に分割する (分割したデータ族の i 番目を $A_i (1 \leq i \leq k)$ とする)．

(3) $\cup_{i\neq 1}^{k} A_i$ のデータに対して先ほどの手法で最小化問題を解く．

(4) 得られた解 $(\beta_{0,1}, \beta_1)$ を用いた A_1 における誤差を求める；A_1 が

$$
\begin{gathered}
(x_{a_1,1}, \cdots, x_{a_1,t}, y_{a_1}) \\
\vdots \\
(x_{a_s,1}, \cdots, x_{a_s,t}, y_{a_s})
\end{gathered}
$$

となっているときに，誤差

$$
e_{s,1} := \left\| \begin{pmatrix} y_{a_1} \\ \vdots \\ y_{a_s} \end{pmatrix} - \begin{pmatrix} \beta_{0,1} \\ \vdots \\ \beta_{0,1} \end{pmatrix} - \begin{pmatrix} x_{a_1,1} & \cdots & x_{a_1,s} \\ \vdots & \ddots & \vdots \\ x_{a_s,1} & \cdots, & x_{a_s,t} \end{pmatrix} \begin{pmatrix} \beta_{1,1} \\ \cdots \\ \beta_{1,s} \end{pmatrix} \right\|^2
$$

を求める．

(5) (3),(4) で $\cup_{i\neq 2} A_i, \cup_{i\neq 3} A_i, \cdots$ の場合にも同様に誤差 $e_{s,2}, e_{s,3}$ を求め，その平均値

$$
e_s := \frac{1}{k} \sum_{i=1}^{N} e_{t,i}
$$

を求める．

(6) 以上を全ての s について行い，e_s が最も小さくなるような s を選ぶ．

p が大きくなるにつれてこのような問題が生じてくる．これらを解決するために，以降では定数 $\lambda \geq 0$ に対して，β の各成分が大きくなるごとに対する罰則を $\|y - X\beta\|^2$ に加えた

$$L := \frac{1}{2N}\|y - X\beta\|^2 + \lambda\|\beta\|_1$$

もしくは

$$L := \frac{1}{N}\|y - X\beta\|^2 + \lambda\|\beta\|_2$$

の値を最小にする β を求める問題を検討する．ここで，$\hat{\beta}$ が求まれば $\hat{\beta}_0$ は求めることができたので，中心化された X, y に対して上式を最小にする $\hat{\beta}$ を求め，それから $\hat{\beta}_0$ を求めることにする．