# 統計解析

古谷知之

## 授業概要

* 履修者の状況に応じて変更される場合がありますが、全体としては以下のような授業構成となります。
* 講義の中でR演習を行うこともあります。

第1回　ガイダンス・単回帰分析
第2回　重回帰分析(1)
第3回　重回帰分析(2)
第4回　一般化線形回帰モデル(1)
第5回　一般化線形回帰モデル(2)
第6回　一般化線形回帰モデル(3)
第7回　一般化線形回帰モデル(4)
第8回　一般化線形回帰モデル(5)
第9回　一般化線形回帰モデル(6)
第10回　一般化線形混合モデル
第11回　状態空間モデル
第12回　R演習(1)
第13回　R演習(2)
第14回　R演習(3)

# 統計モデルの種類

| | 主な推定方法 | データ分布 | 回帰係数 |
|---|---|---|---|
| 線形回帰モデル<br>（単回帰・重回帰など） | 最小二乗法 | 正規分布 | 一変数に一つ |
| 一般化線形モデル | 最尤推定法 | 正規分布以外の分布も可能 | 一変数に一つ |
| 一般化線形混合モデル | 最尤推定法 | 正規分布以外の分布も可能 | 変数の個体差に応じて推定可能 |
| 階層ベイズモデル | ベイズ推定 | 正規分布以外の分布も可能 | 変数の個体差に応じて推定可能 |

# 本授業で扱う統計モデル

- 線形回帰モデル
  - 単回帰モデル、重回帰モデル
- 一般化線形回帰モデル
  - 離散：ポアソン回帰モデル、二項反応モデル（ロジスティック回帰モデル、プロビット回帰モデル、補対数対数モデル）、負の二項分布モデル、ゼロ過剰ポアソン回帰モデル、ゼロ過剰負の二項分布モデル
  - 連続：ガンマ回帰モデル、ベータ回帰モデル、指数-ガウス回帰モデル
  - スパース：Lasso回帰モデル、Ridge回帰モデル
- 一般化線形混合モデル
  - マルチレベルモデル
- 状態空間モデル

# 代表的な一般化線形回帰モデル

- 被説明変数が離散変数
  - 0or1の２値：(二項)ロジスティック回帰モデル、(二項)プロビット回帰モデル、補対数対数モデル
  - ０以上の整数
    - 発生頻度が少ない：ポアソン回帰モデル、負の二項分布モデル
    - 発生頻度０が非常に多い：Hurdleモデル、ゼロ過剰モデル
- 被説明変数が連続変数
  - [0, 1]の確率値：ベータ回帰モデル
  - ０より大きい値：ガンマ回帰モデル、指数-ガウス回帰モデル
- 被説明変数がスパース
  - Lasso回帰モデル、Ridge回帰モデル

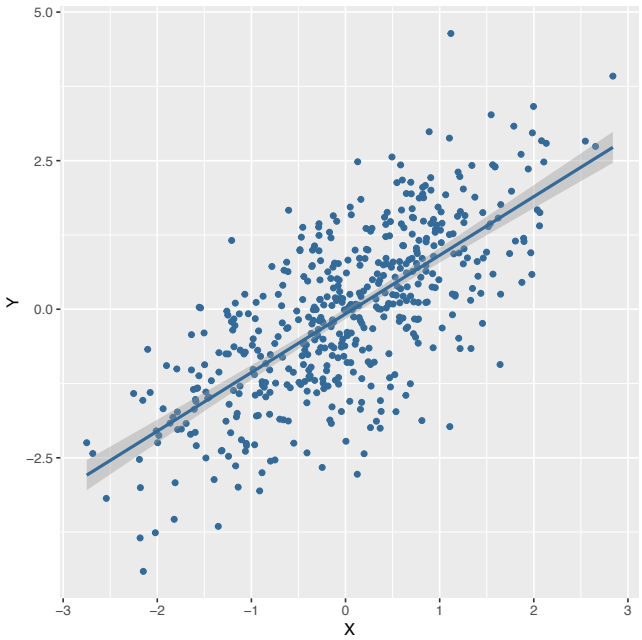
# 授業の内容

- 重回帰モデル（復習）
- 正則化
- Ridge回帰
- Lasso回帰

# 線形回帰モデル

- 説明変数と被説明変数がともに正規分布
- 誤差項も正規分布

- 説明変数と被説明変数との関係が線形式で表される




# 線形回帰モデル（重回帰分析）

- 従属変数$y$と$k$個の独立変数$x_1, x_2, \ldots, x_k$に対する標本数が$n$個の重回帰モデルは以下のように記述できる$(i = 1, \ldots, n)$

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}
$$

# 線形回帰モデル（重回帰分析）

• 次のようなベクトルと行列を用いて、

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

• 次式のように簡略化できる

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

あるいは

$$y_i = X_i^T \boldsymbol{\beta} + \varepsilon_i$$

# 重回帰分析

• 誤差項$\boldsymbol{\varepsilon} = \boldsymbol{y} - X\boldsymbol{\beta}$の二乗和$Q$は、

$$Q = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T X^T X\boldsymbol{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$$

• 最小二乗法より、

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\beta} = 0$$

• ここから以下の正規方程式を得る

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$$

• 両辺に左から$(X^T X)^{-1}$をかけると、回帰係数の推定量$\widehat{\boldsymbol{\beta}}$を得る

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T\boldsymbol{y}$$

# 線形回帰分析を行う上での仮定（前提）

- 線形回帰分析では、独立変数と従属変数がともに正規分布に従うことを前提としている

- 独立変数行列$X$が平均$\mu$、分散$\Sigma$の正規分布に従う$X \sim N(\mu, \Sigma)$とき、$X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\beta}\Sigma\boldsymbol{\beta}^T)$となる

- さらに誤差項$\boldsymbol{\varepsilon}$が平均$0$、分散$\sigma^2$の正規分布に従う$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$と仮定している。
- このことから従属変数$\boldsymbol{y}$は平均$X\boldsymbol{\beta}$、分散$\sigma^2 I$の正規分布に従う

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$$

# 重回帰モデルの統計量

- 回帰係数$\boldsymbol{\beta}$・誤差項$\boldsymbol{\varepsilon}$・従属変数$\boldsymbol{y}$の確率分布から、偏回帰係数$\hat{\boldsymbol{\beta}}$・予測値$\hat{\boldsymbol{y}}$・予測誤差$\boldsymbol{e}$の確率分布は以下のようになる

- 偏回帰係数

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

- 予測値

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \boldsymbol{y} = H\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 H)$$

$$H = X(X^T X)^{-1} X^T$$ $H$はハット行列

- 予測誤差

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - H)\boldsymbol{y} \sim N\big(0, \sigma^2 (I - H)\big)$$

# 重回帰モデルの統計量

- 偏回帰係数は正規分布に従う

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

- この性質を標準化すると、以下のようになる

$$\frac{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\sigma^2 (X^T X)^{-1}}} \sim N(0, 1)$$

# 線形回帰モデルの尤度関数

- データ$X$、未知パラメータ$\boldsymbol{\beta}$、誤差項の分散$\sigma^2$が与えられた条件下で、被説明変数$\boldsymbol{y}$が得られる条件付き確率を尤度関数という
- サンプル$i$の説明変数$x_i = (1, x_{i1}, \ldots, x_{ik})$、被説明変数$y_i$とするとき、尤度関数は以下のような正規分布となる

$$p(y_i | x_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[ -\frac{(y_i - x_i \boldsymbol{\beta})^2}{2\sigma^2} \right]$$

# 尤度関数

- 尤度関数の平均と分散はそれぞれ以下のとおりとなる

- 平均

$$E(y_i|x_i;\boldsymbol{\beta},\sigma^2) = x_i\boldsymbol{\beta}$$

- 分散

$$V(y_i|x_i;\boldsymbol{\beta},\sigma^2) = \sigma^2$$

# 尤度関数

- 全てのサンプル$i$についての尤度関数は

$$p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = \prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\beta},\sigma^2)$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

# 対数尤度関数

- 尤度関数の自然対数をとると

$$\ln[p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2)] = \ln\left[\prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\beta},\sigma^2)\right]$$

$$= \ln\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]\right]$$

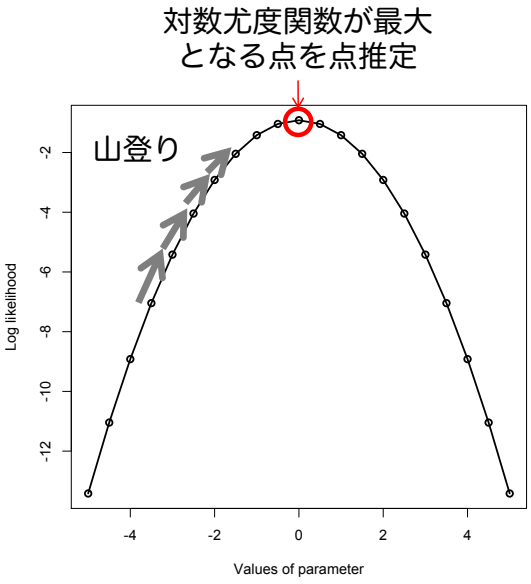
$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$

# 最尤法と最小二乗法

- 最尤法では(対数)尤度関数を最大化することで未知パラメータを得る⇒対数尤度関数は上に凸となる関数

- 対数尤度関数を最大化することは、次式を最小化することと同じ

$$\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$



# 最小二乗法による解

- 次式を最小化することにより得られる未知パラメータはそれぞれ以下のようになる

$$\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$

- 最小二乗解

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\boldsymbol{\beta}}\right)^2$$ 自由度: $n-(k+1)$

# 尤度関数(全データ)

- データ$X$、未知パラメータ$\boldsymbol{\beta}$、分散$\sigma^2$が与えられた条件下で、被説明変数$\boldsymbol{y}$が得られる条件付き確率、すなわち尤度関数は、以下のような正規分布となる

$$p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

# 尤度関数(全データ)

- 尤度関数の平均と分散はそれぞれ以下のとおりとなる

- 平均

$$E(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = X\boldsymbol{\beta}$$

- 分散

$$V(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = \sigma^2$$

# 尤度関数(全データ)

- 全てのデータについての尤度関数は

$$p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}\right]$$

# 対数尤度関数(全データ)

• 全てのデータについての尤度関数は

$$\ln[p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2)] = \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}\right]\right]$$

$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}$$

# 最小二乗法による解(全データ)

• 対数尤度関数を最大化⇔最小二乗法による不偏推定量が得られる

$$\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$$

$$\hat{\sigma}^2 = \frac{1}{\nu}\left(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\right)$$

$\nu = n-(k+1)$ …自由度

↑定数項を加えた変数の数

# 最小二乗法による解が得られる条件

- 最小二乗法により不偏推定量$\widehat{\boldsymbol{\beta}}$を得るためには$(X^T X)^{-1}$が存在しなくてはならない

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

- $X$がフルランクでないとき、$(X^T X)^{-1}$が存在しない
  - $n < k$のとき（サンプル数より説明変数の数が多いとき）
  - 互いに相関する説明変数があるとき
  - 互いに類似するサンプルがあるとき、など

# 高次元データのモデル選択の難しさ

- 「説明変数の数$k$が増える」＝パラメータの次元$k$ が大きい高次元データとなる

- AICでモデル選択をする場合、 $2^k$通りとなる(NP困難)ため、計算に時間がかかる

- そのため、高次元データを用いた回帰モデルはモデル（変数）選択が容易でない

# 行列の行基本変形

- 行列$A$に対して以下の操作を行うことを行基本変形という

1. ある行の定数倍を他の行に足す
2. 行を入れ替える
3. 行に 0 以外のかける

- 行基本変形は、正則行列を左からかけることと対応している

# 階段行列

- 行列$A$を行基本変形することで得られる変形後の行列を階段行列という
- 階段行列は例えば以下のようになる

$$A = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \\ & & \end{pmatrix}$$

# 行列の階数

- 任意の行列$A$を行基本変形により階段行列に変形できる
- 得られた階段行列の、「零ベクトルでない行ベクトル」の個数を行列の階数といい、 $\mathrm{rank}(A)$とあらわす

# 対処方法

- 最小二乗法により解が得られない場合や、高次元データで変数選択を容易にする対処方法として、以下の方法が挙げられる

- 互いに相関が強い説明変数のどちらかを除外する
- 互いに類似する標本のどちらかを除外する
- 擬似逆行列を用いる
- $(X^T X)^{-1}$が求まるようにする→「正則化」という

# 正則化(Regularization)

- $(X^TX)^{-1}$が求まるようにするには、逆行列が求まるように$X^TX$を正則化してやればよい
- 例えば、正規化パラメータ$\lambda$と単位行列$I_{k+1}$を以下のように用いると、正則化することができる

$$\boldsymbol{\beta}_\lambda = (X^TX + \lambda I_{k+1})^{-1}X^T\boldsymbol{y}$$

- 上式の解を得るには、以下の誤差関数$R_\lambda(\boldsymbol{\beta})$を最小化するような$\boldsymbol{\beta}$を推定すれば良い

$$R_\lambda(\boldsymbol{\beta}) = (y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

# 正則行列

- 一般に、以下の性質が成り立つ$n\times n$の正方行列$A$を正則行列という

$$AB = I_n = BA$$

- ここで、$I_n$は$n\times n$の単位行列、$B$は$n\times n$の正方行列

# $L_p$正則化

- 以下の重回帰モデルが与えられたとき

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- 以下のような誤差関数を、 $L_p$正則化を導入した誤差関数という

$$R_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_q = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\sum_{j=1}^{k}\left|\beta_j\right|^q$$

- $\lambda\|\boldsymbol{\beta}\|_q = \lambda\sum_{j=1}^{k}\left|\beta_j\right|^q$はペナルティ項という（$q \geq 0, \lambda \geq 0$）
- $\lambda$は正則化係数 complexity parameter。小さいと（変数が多くなり）複雑なモデル、大きいと簡素なモデルとなる
- 誤差関数$R_\lambda(\boldsymbol{\beta})$を最小化することで、 $\boldsymbol{\beta}$が求まる

# $L_p$正則化

- $L_p$正則化を導入した誤差関数$R_\lambda(\boldsymbol{\beta})$を最小化する$\boldsymbol{\beta}$の集合を求めるには、次式を計算すれば良い

$$\begin{aligned}
&\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} R_\lambda(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}\left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_q\right] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}\left[(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\sum_{j=1}^{k}\left|\beta_j\right|^q\right] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}\left|\beta_j\right|^q\right]
\end{aligned}$$

# arg minとは

- $\max f(x)$： $f(x)$の最大値
- $\arg\max f(x)$： $f(x)$を最大にする$x$の集合

- $\min f(x)$： $f(x)$の最小値
- $\arg\min f(x)$： $f(x)$を最小にする$x$の集合

- 定理：下に凸な関数の$\arg\min$は凸集合

# $L_q$正則化

- $L_q$正則化を導入した誤差関数$R_\lambda(\boldsymbol{\beta})$を最小化するとき、

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} R_\lambda(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \left[\|y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_q\right]$$

- $q = 2$のときRidge回帰モデル
- $q = 1$のときLasso回帰モデル
- また$0 < \alpha < 1$となる$\alpha$を用いてペナルティ項$\lambda\|\boldsymbol{\beta}\|_q$を以下のように置き換えるとき、Elastic-net (EN) と呼ばれる($\alpha$を指定)

$$\lambda\sum_{j=1}^{k}\{2\alpha|\beta_j| + (1-\alpha)\beta_j^2\}$$

# Ridge、Lasso、ENの違い

- $L_q$正則化を導入した誤差関数$R_\lambda(\boldsymbol{\beta})$の最小化問題

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} R_\lambda(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \left[\|y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_q\right]$$

- Ridge回帰モデル$(q = 2)$
  - 誤差の二乗和＋回帰係数の二乗和を最小化
- Lasso回帰モデル$(q = 1)$
  - 誤差の二乗和＋回帰係数の絶対値の和を最小化
- Elastic-net$(0 < \alpha < 1)$
  - 誤差の二乗和＋回帰係数の二乗和＋回帰係数の絶対値の和を最小化

# $L_2$正則化とRidge回帰

- $q = 2$のとき誤差関数$R_\lambda(\boldsymbol{\beta})$は以下のようになる

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} R_\lambda(\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} [\|y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \left[(y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta}) + \lambda\sum_{j=1}^{k}|\beta_j|^2\right] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}|\beta_j|^2\right]
\end{aligned}$$

# $L_2$正則化とRidge回帰

- Ridge回帰の回帰係数$\hat{\boldsymbol{\beta}}^{\text{ridge}}$を得る問題は、以下の最適化問題と等価である

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min} \left[ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right]$$

← OLSの誤差項

$$s.t. \sum_{j=1}^{k} {\beta_j}^2 \leq t$$

← 制約条件

# $L_2$正則化とRidge回帰

- これは以下のような行列式としても表せる

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min} R_\lambda(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min} [(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}]$$

- すると、Ridge回帰のパラメータは次式により求められる

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (X^T X + \lambda I_{k+1})^{-1} X^T \boldsymbol{y}$$

# $L_2$正則化とRidge回帰

- 証明

$$R_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2$$
$$= (\boldsymbol{y} - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

- これを$\boldsymbol{\beta}$で微分すると

$$\frac{R_\lambda(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta}$$

- $\frac{R_\lambda(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0$より

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (X^TX + \lambda I_{k+1})^{-1}X^T\boldsymbol{y}$$

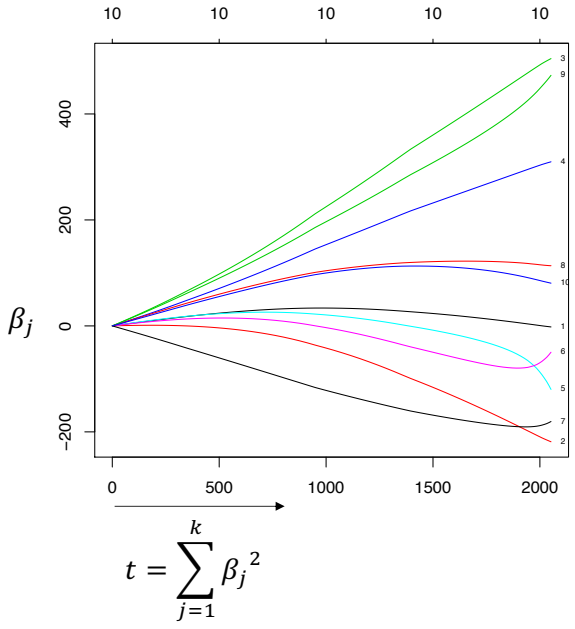
# $L_2$正則化とRidge回帰

- $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$を得る最適化問題

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right]$$

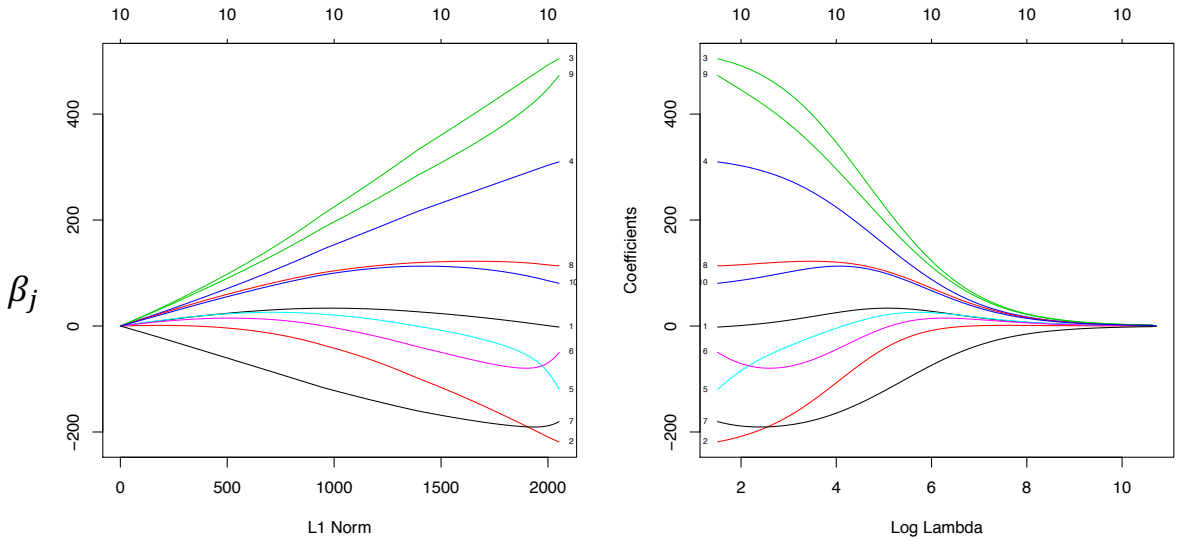
$$s.t.\sum_{j=1}^{k}{\beta_j}^2 \le t$$

- $t$を0から漸次大きくすると$\beta_j$は右図のようになる
- $L_2$ノルムが小さい($\lambda$が大きい)とき、推定量はスパースとなる



# $L_2$正則化とRidge回帰

- $L_2$ノルムが小さい($\lambda$ が大きい)とき、推定量はスパースとなる

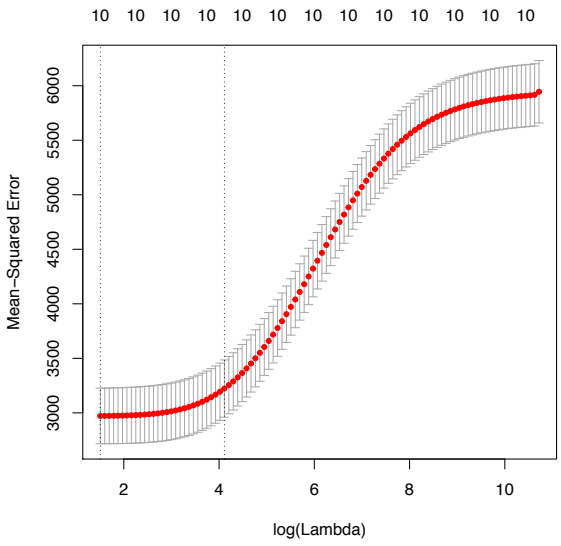



# $L_2$正則化とRidge回帰

- 誤差関数$R_\lambda(\boldsymbol{\beta})$

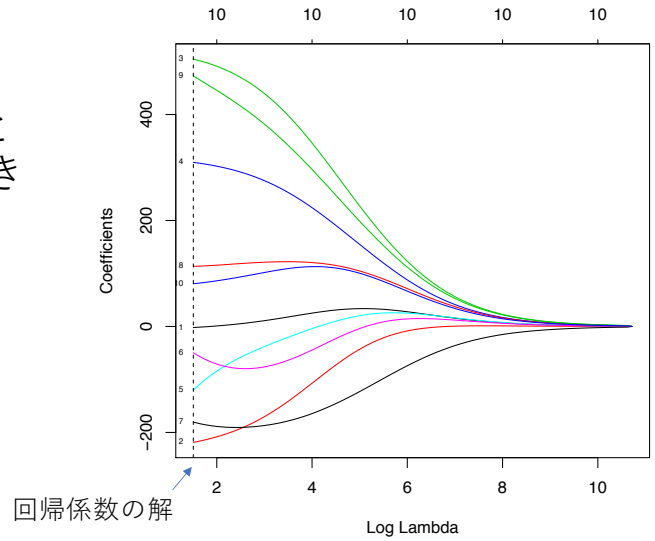
$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}[\|y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2]$$

- $\lambda$はクロスバリデーション (CV) の MSE（平均二乗誤差）が最小となるような値を用いる



# $L_2$正則化とRidge回帰

- クロスバリデーション（CV）のMSE（平均二乗誤差）が最小となる$\lambda$の回帰係数$\beta_j$が求めるべき回帰係数




# $L_1$正則化とLasso回帰

- $q = 1$のとき誤差関数$R_\lambda(\boldsymbol{\beta})$は以下のようになる

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\text{lasso}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} R_\lambda(\boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} [\|y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \left[ (y - X\boldsymbol{\beta})^T (y - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{k} |\beta_j| \right] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \left[ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right]
\end{aligned}$$

# $L_1$正則化とLasso回帰

- Lasso回帰の回帰係数$\widehat{\boldsymbol{\beta}}^{\text{lasso}}$を得る問題は、以下の最適化問題と等価である

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min} \left[ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right]$$
$$s.t. \sum_{j=1}^{k} |\beta_j| \leq t$$

# $L_1$正則化とLasso回帰

- Lasso回帰の回帰係数$\widehat{\boldsymbol{\beta}}^{\text{lasso}}$を得る最適化問題

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min} \left[ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right]$$
$$s.t. \sum_{j=1}^{k} |\beta_j| \leq t$$

- 制約変数$t \geq 0$
- 最小二乗法による回帰モデルの不偏推定量を$\hat{\beta}$とし、 $t_0 = \sum_{j=1}^{k} \hat{\beta}_j$とすると、 $t < t_0$

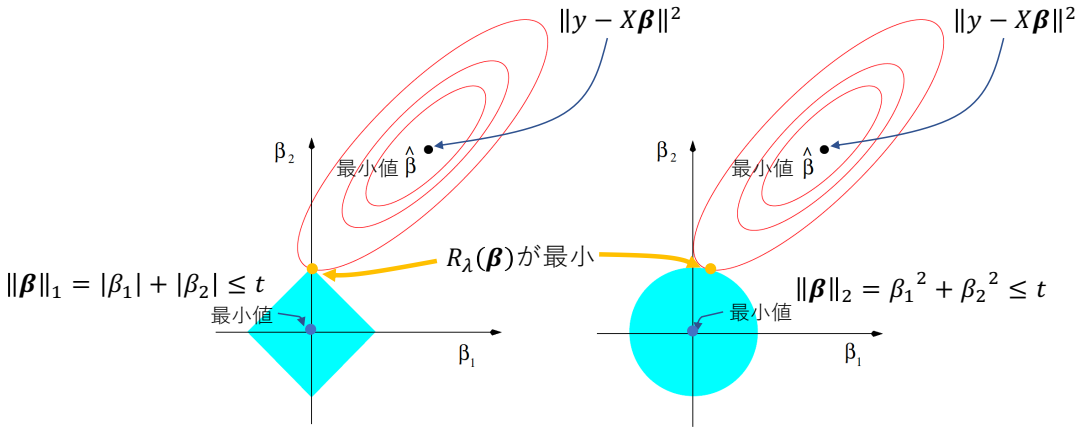
# Lasso回帰とRidge回帰

- Ridge回帰とLasso回帰の最適化問題は、$\|\boldsymbol{\beta}\|_q$の制約条件の下で回帰モデルの誤差項の二乗和について最適な組み合わせを求めることにほかならない

- いま2つのパラメータ$\beta_1$と$\beta_2$のみを考えると、$\|\boldsymbol{\beta}\|_q$の制約条件はそれぞれ次式のようになる
- Lasso回帰：$|\beta_1| + |\beta_2| \leq t$（回帰係数の絶対値の和）
- Ridge回帰：$\beta_1{}^2 + \beta_2{}^2 \leq t$（回帰係数の二乗和）

# Lasso回帰とRidge回帰

- 最小二乗法による回帰モデルの不偏推定量を$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$とする
- $\hat{\beta}$および$\beta_1$と$\beta_2$の制約条件との関係を誤差空間の中で視覚的に表現すれば、以下のようになる

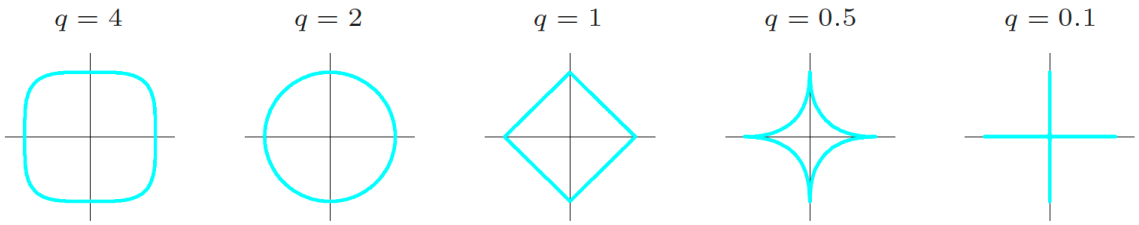



Lasso回帰　　Ridge回帰

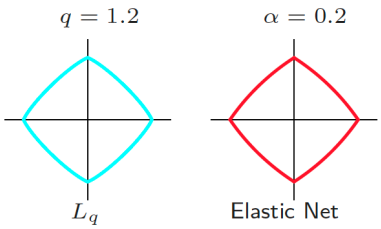
Hastie et. al. (2017)

# $\|\boldsymbol{\beta}\|_q$の挙動

- ペナルティ項に含まれる$\|\boldsymbol{\beta}\|_q = \sum_{j=1}^{k}|\beta_j|^q$の境界値は以下のようになる




- またElastic-net $\lambda\sum_{j=1}^{k}\{(1-\alpha)\beta_j^2 + 2|\beta_j|\}$は以下のようになる




Hastie et. al. (2017)

# $L_1$正則化とLasso回帰

- なぜ$L_1$正則化をするとよいのか？

- $L_1$正則化を行うと回帰係数$\boldsymbol{\beta}$の一部が 0 になりスパースな解が得られる→説明変数が多い場合に変数選択をしてくれる

- 従来はモデル推定とモデル選択は別々に行っていたが、同時に行える

# $L_1$正則化とLasso回帰

- Lasso回帰の回帰係数$\widehat{\boldsymbol{\beta}}^{\mathbf{lasso}}$を得る最適化問題は以下の最適化問題と同じ問題に帰着できる

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \frac{1}{2}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right]$$
$$s.t. \sum_{j=1}^{k}|\beta_j| \leq t$$

- ここで、 $x_{ij}$は平均0、分散1に標準化されているとする

# $L_1$正則化とLasso回帰

- 従って、 Lasso回帰の回帰係数$\widehat{\boldsymbol{\beta}}^{\mathbf{lasso}}$を得る最適化問題は、以下のラグランジュ関数$f(\beta)$を最小化する問題と等価となる

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \gamma\sum_{j=1}^{k}|\beta_j|$$

ただし、 $\gamma$ ($\gamma \geq 0$ )はラグランジュ係数

# $L_1$正則化とLasso回帰

- 最小二乗法による回帰モデルの不偏推定量を$\hat{\beta}$とする
- このとき、 ラグランジュ関数$f(\beta)$の最小解$\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}(\gamma)$は$\hat{\beta}$と$\gamma$の関数を用いて以下のように求められる

$$\widehat{\boldsymbol{\beta}}^{\mathrm{lasso}}(\gamma) = \mathrm{sin}(\hat{\beta})\left(|\hat{\beta}| - \gamma\right)_+$$

# $L_1$正則化とLasso回帰

- $\lambda\|\boldsymbol{\beta}\|_1$は$\boldsymbol{\beta}$で微分できない点を含むため、以下のようなアルゴリズムを用いて回帰係数$\boldsymbol{\beta}$を求める方法が提案されている

- Coordinated Descent Algorithm
  - https://www.jstatsoft.org/article/view/v033i01
- Least Angle Regression (LARS)
  - https://projecteuclid.org/euclid.aos/1083178935
- Iterative Shrinkage Thresholding Algorism (ISTA)
  - https://people.rennes.inria.fr/Cedric.Herzet/Cedric.Herzet/Sparse_Seminar/Entrees/2012/11/12_A_Fast_Iterative_Shrinkage-Thresholding_Algorithmfor_Linear_Inverse_Problems_(A._Beck,_M._Teboulle)_files/Breck_2009.pdf
- Alternating Direction Method of Multipliers (ADMM)
  - https://web.stanford.edu/~boyd/papers/pdf/admm_distr_stats.pdf

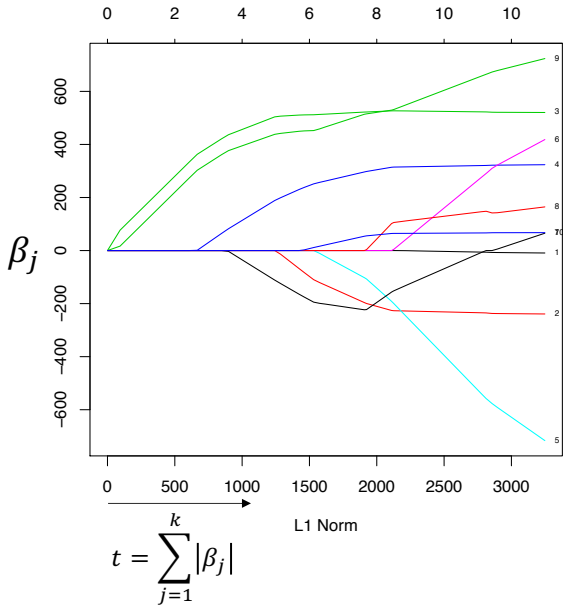
# $L_1$正則化とLasso回帰

- Lasso回帰の回帰係数$\hat{\boldsymbol{\beta}}^{\text{lasso}}$を得る最適化問題

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \frac{1}{2}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right]$$

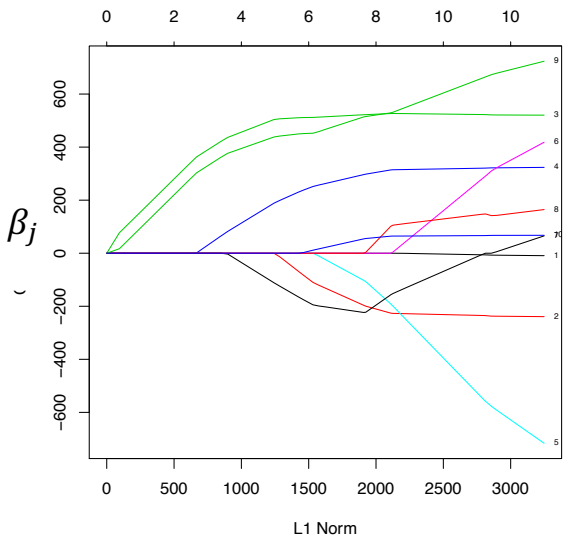
$$s.t. \sum_{j=1}^{k}|\beta_j| \le t$$

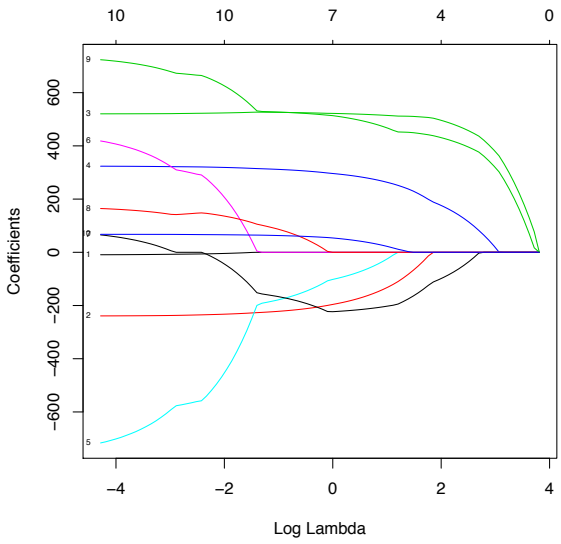
- $t$を0から漸次大きくすると$\beta_j$は右図のようになる
- $L_1$ノルムが小さい($\lambda$が大きい)とき、推定量はスパースとなる




# $L_1$正則化とLasso回帰

- $L_1$ノルムが小さい($\lambda$が大きい)とき、推定量はスパースとなる



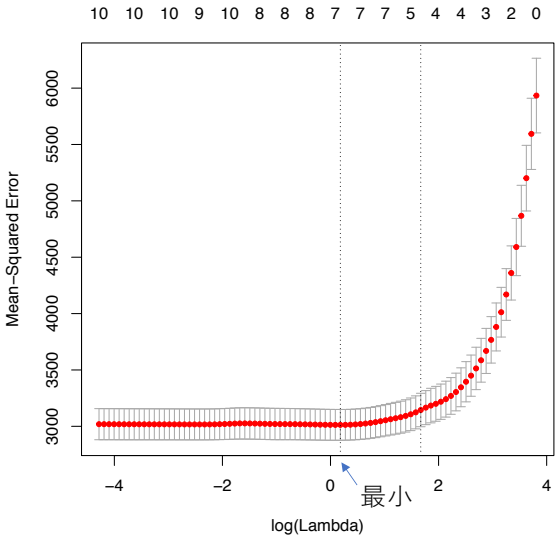



# $L_1$正則化とLasso回帰

- 誤差関数$R_\lambda(\boldsymbol{\beta})$

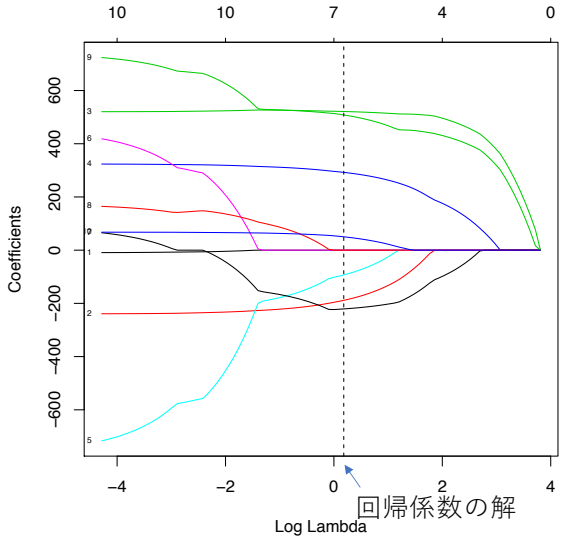
$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min}[\|y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1]$$

- $\lambda$はクロスバリデーション (CV) の MSE（平均二乗誤差）が最小となるような値を用いる




# $L_1$正則化とLasso回帰

- クロスバリデーション (CV) の MSE（平均二乗誤差）が最小となる$\lambda$の回帰係数$\beta_j$が求めるべき回帰係数



# 分析の適用例

- Rのlarsパッケージにある”diabetes”データ
  - https://web.stanford.edu/~hastie/StatLearnSparsity_files/DATA/diabetes.html
- 442名の糖尿病患者データ
- 基礎項目（標準化済み）：age, sex, BMI (Body mass index), map (平均動脈圧)
- 血清検査測定項目(x)：
  - tc (トリグリセライド？) (総コレステロール)
  - ldl (LDLコレステロール：低密度リポタンパク質=悪玉コレステロール)
  - hdl (HDLコレステロール：高密度リポタンパク質=善玉コレステロール)
  - tch (総コレステロール)
  - ltg (中性脂肪：トリグリセリド)
  - glu (血糖値：グルコース)
- ターゲット項目(y)：１年後の進行状況
- 行列X：標準化されたデータ