04. 重回帰分析

京都大学　　加納　学

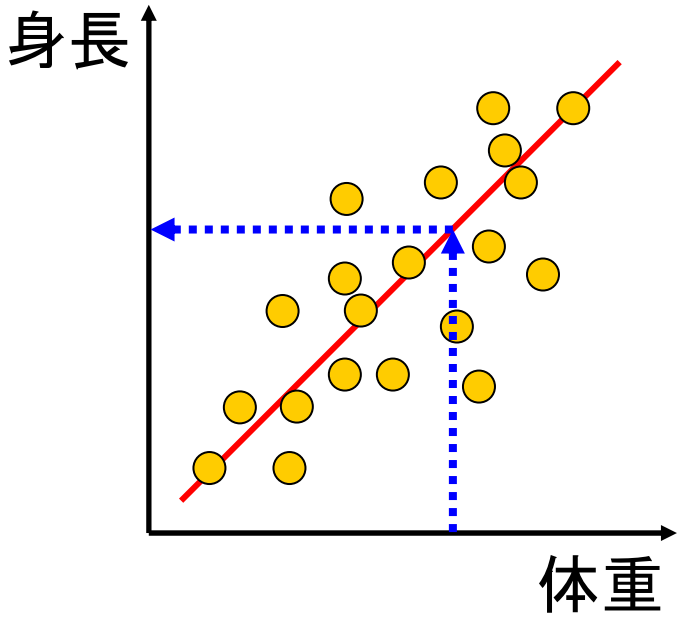
Division of Process Control & Process Systems Engineering
Department of Chemical Engineering, Kyoto University

manabu@cheme.kyoto-u.ac.jp
http://www-pse.cheme.kyoto-u.ac.jp/~kano/

Outline

- 重回帰式の導出
- 幾何学的解釈
- 重回帰式の評価
- 具体例
- 多重共線性
- リッジ回帰

身近な例



体重から身長を推定できる？

身長 = 定数 × 体重 + 定数 + 誤差
y　　b_1　　x　　b_0　　e

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_1 x_i - b_0)^2$$ を最小にする係数 b を求める.

因果関係を探る

- 重回帰分析
  結果である目的変数 y と原因である説明変数 $\{x_p | p = 1, 2, \cdots, P\}$ の関係を重回帰式で表現する手法

$$y = \sum_{p=1}^{P} b_p x_p + b_0$$

（標本）偏回帰係数

- 現実には，目的変数は説明変数以外の要因にも影響されるため，それらの n 番目の標本(測定値)が単回帰モデルによって表現されると考える.

$$y_n = \sum_{p=1}^{P} \beta_p x_{np} + \beta_0 + \varepsilon_n$$

母偏回帰係数

誤差項 ε_n は互いに独立に $N(0, \sigma^2)$ に従うと仮定する.

目的変数の予測

- 目的変数の予測値
  - 各変数の平均を 0 とすれば $b_0 = 0$
  - 誤差項 ε_n の期待値は 0

$$y = \sum_{p=1}^{P} b_p x_p$$

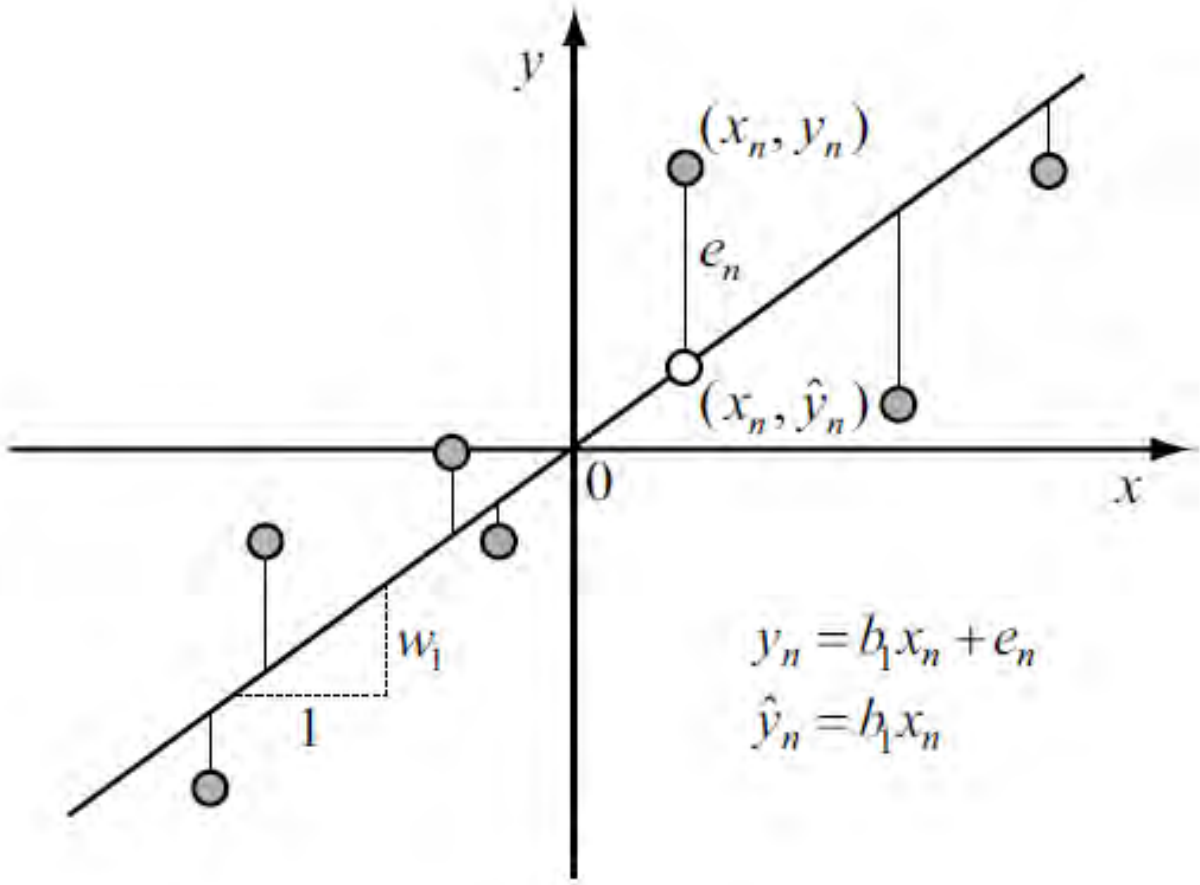
- 残差

  目的変数の測定値と予測値の差

$$\hat{y}_n = \sum_{p=1}^{P} b_p x_{np}$$

回帰分析における誤差の考え方

- 目的変数 y に影響を与える説明変数 x 以外の要因をまとめて誤差とみなすため，y のみに誤差がある，つまり，x は正確に指定できると考える．



最小二乗法

- 最小二乗法
  残差平方和（目的変数の測定値と推定値の差の二乗和）が最小となるように，偏回帰係数を決定する．

予測値

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$$

残差平方和

$$Q = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

$$= (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}})$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NP} \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_P \end{bmatrix}$$

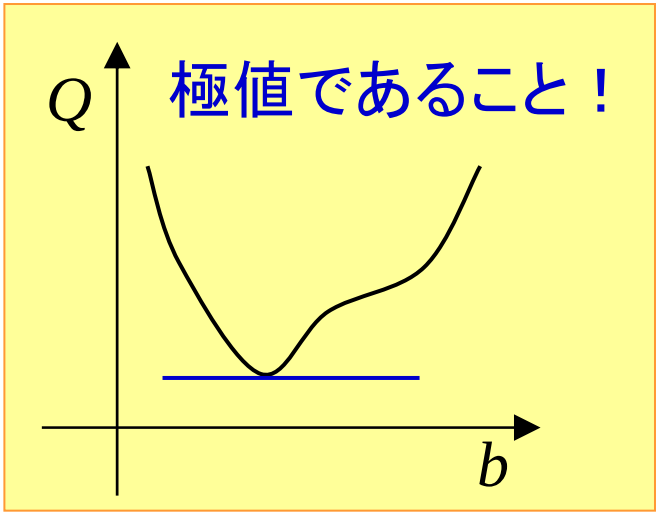
正規方程式の導出

- 残差平方和

$$Q = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}})$$
$$= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b}$$

- 必要条件

$$\frac{1}{2}\frac{\partial Q}{\partial \boldsymbol{b}} = \frac{1}{2}\left[\frac{\partial Q}{\partial b_1} \quad \frac{\partial Q}{\partial b_2} \quad \cdots \quad \frac{\partial Q}{\partial b_P}\right]^T = \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b} - \boldsymbol{X}^T \boldsymbol{y} = 0$$



正規方程式

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

偏回帰係数の推定

- 正規方程式

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

$$\boldsymbol{V}_{XX}\boldsymbol{b} = \boldsymbol{V}_{Xy}$$

- 偏回帰係数の推定値

行列 $\boldsymbol{X}^T\boldsymbol{X}$ が正則である（逆行列を持つ）場合

$$\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

共分散行列

$$\boldsymbol{V}_{XX} = \frac{1}{N-1}\boldsymbol{X}^T\boldsymbol{X} \qquad \boldsymbol{V}_{Xy} = \frac{1}{N-1}\boldsymbol{X}^T\boldsymbol{y}$$

標準化

各変数を平均0，分散1の変数に変換する．

$$x_{nm} = \frac{x^*_{nm} - \bar{x}_m}{\sigma_m}$$

変数　m
サンプル　n

平均

$$\bar{x}_m = \frac{1}{N}\sum_{n=1}^{N} x^*_{nm}$$

分散

$$\sigma_m^2 = \frac{1}{N-1}\sum_{n=1}^{N} (x^*_{nm} - \bar{x}_m)^2$$

重回帰分析： 重回帰式

標準化後の変数による表現

$$\hat{y} = \sum_{p=1}^{P} b_p x_p$$

b_p 標準偏回帰係数

標準化前の変数による表現

$$\frac{\hat{y}^* - \bar{y}}{\sigma_y} = \sum_{p=1}^{P} b_p \frac{x_p^* - \bar{x}_p}{\sigma_p}$$

$\frac{\sigma_y}{\sigma_p} b_p$ 偏回帰係数

$$\hat{y}^* = \sum_{p=1}^{P} \frac{b_p \sigma_y}{\sigma_p} x_p^* + \left(\bar{y} - \sum_{p=1}^{P} \frac{b_p \sigma_y}{\sigma_p} \bar{x}_p \right)$$

Outline

- 重回帰式の導出
- 幾何学的解釈
- 重回帰式の評価
- 具体例
- 多重共線性
- リッジ回帰

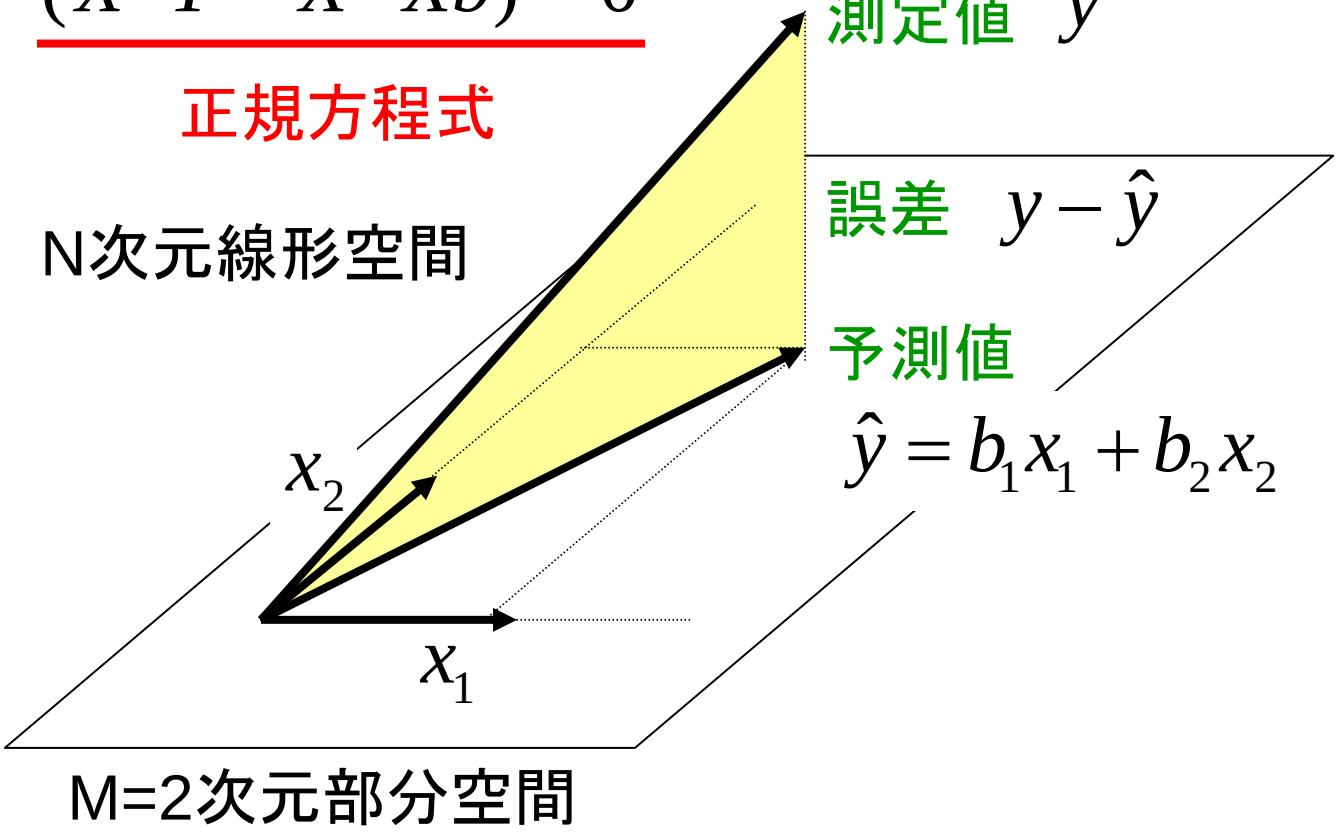
重回帰分析の幾何学的解釈

誤差が最小となるためには，誤差と予測値が直交すればよい．

$$\langle \hat{y}, y-\hat{y} \rangle = \langle Xb, Y-Xb \rangle$$

$$= b^T (X^T Y - X^T Xb) = 0$$

正規方程式



重相関係数の最大化

誤差が最小となるためには，誤差と予測値が直交すればよい．

誤差が最小となるためには，
測定値と予測値がなす角 θ が最小になればよい．

誤差が最小となるためには，
測定値と予測値の相関係数が最大になればよい．

重相関係数　$$r_{y\hat{y}} = \frac{s_{y\hat{y}}^2}{s_y s_{\hat{y}}} = \frac{y^T \hat{y}}{\|y\|\|\hat{y}\|} = \cos\theta$$

重回帰式の評価

- **重相関係数**
  目的変数 y とその推定値 $\hat{y}$ の相関係数

$$R = \frac{s_{y\hat{y}}^2}{s_y s_{\hat{y}}} = \frac{\langle \boldsymbol{y}, \hat{\boldsymbol{y}} \rangle}{\|\boldsymbol{y}\| \|\hat{\boldsymbol{y}}\|} = \cos\theta$$

- **寄与率（決定係数）**
  目的変数 y の分散に対する推定値 $\hat{y}$ の分散の比

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \left(\frac{s_{y\hat{y}}^2}{s_y s_{\hat{y}}} \right)^2 = r_{y\hat{y}}^2$$

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

寄与率に関する式

$$s_y^2 = \frac{1}{N-1}\langle \boldsymbol{y}, \boldsymbol{y} \rangle = \frac{1}{N-1}\|\boldsymbol{y}\|^2$$

$$s_{\hat{y}}^2 = \frac{1}{N-1}\langle \hat{\boldsymbol{y}}, \hat{\boldsymbol{y}} \rangle = \frac{1}{N-1}\|\hat{\boldsymbol{y}}\|^2$$

$$s_{y\hat{y}}^2 = \frac{1}{N-1}\langle \boldsymbol{y}, \hat{\boldsymbol{y}} \rangle = \frac{1}{N-1}\|\boldsymbol{y}\|\|\hat{\boldsymbol{y}}\|\cos\theta = \frac{1}{N-1}\|\hat{\boldsymbol{y}}\|^2$$

$$s_{\hat{y}}^2 = s_{y\hat{y}}^2$$

$$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{y}}\|^2 + \|\boldsymbol{e}\|^2$$

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

Outline

- 重回帰式の導出
- 幾何学的解釈
- 重回帰式の評価
- 具体例
- 多重共線性
- リッジ回帰

分散分析

変動要因	平方和	自由度	不偏分散	分散比
全変動	SS_y	$N-1$	–	–
回帰による変動	SS_r	P	$V_r = \frac{SS_r}{P}$	$F = \frac{V_r}{V_e}$
残差の変動	SS_e	$N-P-1$	$V_e = \frac{SS_e}{N-P-1}$	

分散比F は自由度 P, N-P-1 の F 分布に従う.

$F > F(P, N-P-1; \alpha)$ であれば，重回帰式は無意味ではない.

自由度 P, N-P-1 の F 分布，危険率α

分散分析の心

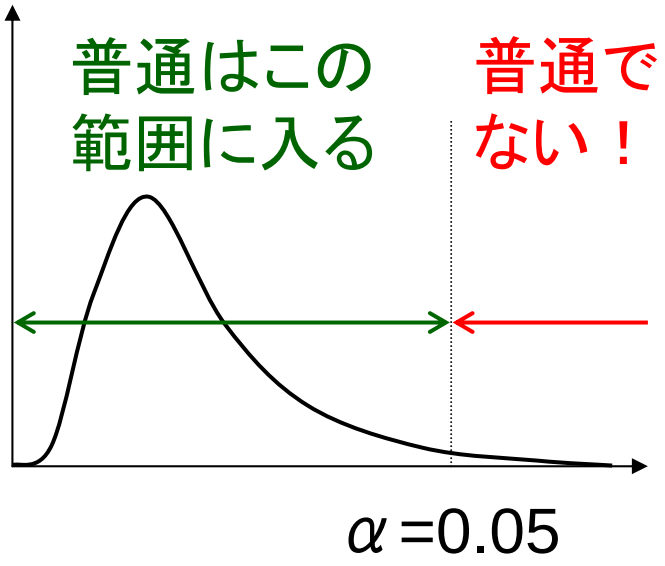
分散比F は自由度 $P, N\text{-}P\text{-}1$ の F 分布に従う.

$F > F(P, N-P-1; \alpha)$ であれば，重回帰式は無意味ではない.

自由度 $P, N\text{-}P\text{-}1$ の F 分布，危険率α

でたらめに重回帰式を作ったとしよう.
そのとき，分散比F はあるF 分布に従う.

もし，F が普通でないほど大きかったら，
つまり，回帰による変動が残差の変動を凌駕していれば，
その重回帰式は無意味ではない！



重要な式

$$y^* - \bar{y} = \sum_{p=1}^{P} b_p (x_p^* - \bar{x}_p)$$

$$F = \frac{V_r}{V_e} = \frac{R^2 / p}{(1-R^2)/(N-p-1)}$$

$$SS_y = \sum_{i=1}^{N} (y_i^* - \bar{y})^2$$

$$SS_r = \sum_{i=1}^{N} (\hat{y}_i^* - \bar{y})^2$$

$$SS_e = \sum_{i=1}^{N} (y_i^* - \hat{y}_i)^2$$

$$SS_y = SS_r + SS_e$$

F分布表（α=0.05）

自由度1

自由度2	1	2	3	4	5	6	7	8
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510

Outline

- 重回帰式の導出
- 幾何学的解釈
- 重回帰式の評価
- 具体例
- 多重共線性
- リッジ回帰

例： 対象データ

	身長(y)	胸囲(x1)	体重(x2)
1	167.0	84.0	61.0
2	167.5	87.0	55.5
3	168.4	86.0	57.0
4	172.0	85.0	57.0
5	155.3	82.0	50.0
6	151.4	87.0	50.0
7	163.0	92.0	66.5
8	174.0	94.0	65.0
9	168.0	88.0	60.5
10	160.4	84.9	49.5

例： 重回帰分析

	身長(y)	胸囲(x1)	体重(x2)
平均	164.7	87.0	57.2
標準偏差	7.18	3.63	6.13
偏回帰係数	—	-0.427	0.969
標準偏回帰係数	—	-0.216	0.828
重相関係数（R）	0.687	—	—
決定係数（R2）	0.472	—	—

例： 分散分析

変動要因	平方和	自由度	不偏分散	分散比
全変動	464.1	9	—	—
回帰による変動	219.0	2	109.5	3.13
残差の変動	245.1	7	35.0	

$F(P, N-P-1; \alpha)$　自由度 P, N-P-1 の F 分布，危険率α

$F(2,7;0.05) = 4.737 > 3.13$　重回帰式に意味なし！

Outline

- 重回帰式の導出
- 幾何学的解釈
- 重回帰式の評価
- 具体例
- 多重共線性
- リッジ回帰

重回帰分析の問題点

偏回帰係数　$b = (X^T X)^{-1} X^T Y$

$X^T X$ が逆行列を持たない場合，最小二乗法は使えない．

入力変数が線形従属である場合

サンプル数が入力変数の数より少ない場合もダメ．
以下では，サンプル数は十分にあるとする．

多重共線性

	Data “A”			Data “B”		
y	x1	x2	x3	x1	x2	x3
241	15.9	34.6	64.8	16.1	34.7	65.1
321	37.0	16.1	72.1	36.9	16.3	72.0
82	61.1	83.0	28.6	60.6	82.8	28.9
156	86.0	65.9	33.9	85.9	65.9	34.2
係数						

入力変数が厳密に線形従属でなくても，入力変数間に強い相関関係が存在する場合には，係数推定値の分散が大きくなり，推定結果の信頼性が低下してしまう.

何が問題なのか？

推定値の分散が大きくなると，何が問題なのか？
推定ができれば良いのではないか？

＜重回帰分析で酷い目に遭う例＞

$$y = a_1 x_1 + a_2 x_2 \qquad y = x_1 = x_2$$

測定データ　$y = 1.00,\ x_1 = 1.01,\ x_2 = 0.99$

Model 1	$\hat{y} = x_2$	0.99
Model 2	$\hat{y} = 0.5x_1 + 0.5x_2$	1.00
Model 3	$\hat{y} = 100x_1 - 99x_2$	2.99

係数が大きいほど，測定ノイズの影響を受けやすい．

最小二乗法の拡張

Ordinary Least Squares (OLS)

$$a = (X^T X)^{-1} X^T Y \qquad \min \|Y - Xa\|^2$$

Minimum Norm Solution

$$a = X^+ Y \qquad X^+ : \text{一般化逆行列}$$

Ridge Regression (RR)

$$a = (X^T X + \lambda I)^{-1} X^T Y \qquad \min \|Y - Xa\|^2 + \lambda \|a\|^2$$

Principal Component Regression (PCR)
Partial Least Squares (PLS)

いずれの手法も係数を小さく抑えようとする.

Outline

- 重回帰式の導出
- 幾何学的解釈
- 重回帰式の評価
- 具体例
- 多重共線性
- リッジ回帰

リッジ回帰

評価関数の違い

重回帰　$\min \|Y - Xa\|^2$

リッジ回帰　$\min \|Y - Xa\|^2 + \lambda \|a\|^2$

回帰係数に対する懲罰

必要条件（評価が最小となるための）

$$\frac{\partial J}{\partial a} = 2(X^T Xa - X^T Y + \lambda a) = 0$$

$$a = (X^T X + \lambda I)^{-1} X^T Y$$

例題： リッジ回帰

		Data Set: A			Data Set: B		
	y	x1	x2	x3	x1	x2	x3
1	241	15.9	34.6	64.8	16.1	34.7	65.1
2	321	37.0	16.1	72.1	36.9	16.3	72.0
3	82	61.1	83.0	28.6	60.6	82.8	28.9
4	156	86.0	65.9	33.9	85.9	65.9	34.2
偏回帰係数		—	—	—	—	—	—
重回帰		1.36	-0.80	5.01	-4.28	-18.9	-26.0
リッジ回帰		0.86	-2.34	2.36	0.87	-2.38	2.34