応用システム工学
第四回　回帰分析

平成25年05月31日
直線回帰
線形重回帰

相関解析
相関係数

- 二つの確率変数の間の類似の度合い(相関関係の強さ)を示す統計学的な指標
  - 二組の数値からなるデータ列

$$(x, y) = \{(x_i, y_i)\}, i = 1, \cdots, n$$

  - 相関係数ρ_{xy} $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ $\bar{x}, \bar{y}$

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

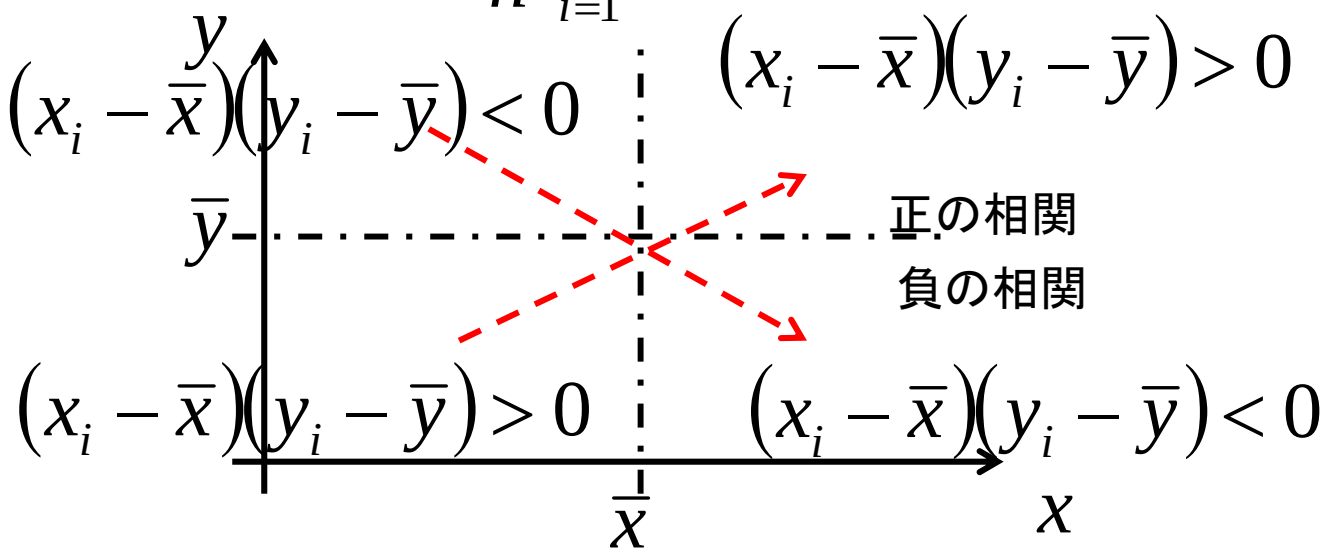
は各々の相加平均

$$|\rho_{xy}| < 1$$

共分散

- 共分散
  - 二組の対応するデータ間での，平均からの偏差の積の平均値

$$\mathrm{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$



回帰分析とは

- データ
  - 目的変数y
  - 目的変数に影響をおよぼす説明変数x
- 分析
  - 予測式を求める
    （ある変量の変化をもう一方の変量の変化で説明するための関数を求める）

$$a_0 + a_1 x_1 + \cdots + a_p x_p \rightarrow y$$

線形回帰（直線回帰）

- 説明変数の数による回帰分析の分類
  - 1個　→　単回帰分析
  - 2個以上　→　重回帰分析

n個のデータの
目的変数yと
説明変数xの組

例
地区数n
地区iの
世帯数x_i，ごみの量y_i

	目的変数:y	説明変数:x
1	y_1	x_1
2	y_2	x_2
…		
i	y_i	x_i
…		
n	y_n	x_n

線形回帰モデル

- モデル式

$$y_i = a_0 + a_1 x_i + e_i \quad (i = 1,2,\cdots,n)$$

  - 未知の定数 a_0, a_1
  - 予測誤差 e_i
- 予測誤差が最小となる定数 $\hat{a}_0, \hat{a}_1$ を求める
  - 予測誤差の平方和を最小にする

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left\{ y_i - \left(a_0 + a_1 x_i \right) \right\}^2$$

    - 最小二乗法→ $\hat{a}_0, \hat{a}_1$

線形回帰モデルのパラメータ同定

- 説明変数x_i，目的変数y_iの平均，分散，共分散を用いて $\hat{a}_0, \hat{a}_1$ を表す
  - 平均
    - 説明変数 $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
    - 目的変数 $$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

線形回帰モデルのパラメータ同定

– 分散　$s_{xx} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i{}^2 - 2x_i\bar{x} + \bar{x}^2\right)$

- 説明変数

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i{}^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i{}^2 - 2\bar{x}n\bar{x} + n\bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \bar{x}^2$$

- 目的変数

$$s_{yy} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i{}^2 - \bar{y}^2$$

線形回帰モデルのパラメータ同定

– 共分散（説明変数と目的変数）

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - x_i\bar{y} - \bar{x}y_i + \overline{xy}\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \overline{xy}\sum_{i=1}^{n} 1\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{y}n\bar{x} - \bar{x}n\bar{y} + \overline{xy}n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

- 同様に $s_{xy} = s_{yx}$

予測誤差平方和の最小条件

- 予測誤差の平方和を a_0, a_1 の関数表現

$$F(a_0, a_1) = \sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_i)\}^2$$

$$= \sum_{i=1}^{n} \left\{ {y_i}^2 - 2y_i(a_0 + a_1 x_i) + (a_0 + a_1 x_i)^2 \right\}$$

$$= \sum_{i=1}^{n} \left\{ {a_0}^2 + (x_i a_1)^2 - 2y_i a_0 - 2y_i x_i a_1 + 2x_i a_0 a_1 + {y_i}^2 \right\}$$

$$= n{a_0}^2 + {a_1}^2 \sum_{i=1}^{n} {x_i}^2 - 2a_0 \sum_{i=1}^{n} y_i - 2a_1 \sum_{i=1}^{n} y_i x_i + 2a_0 a_1 \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} {y_i}^2$$

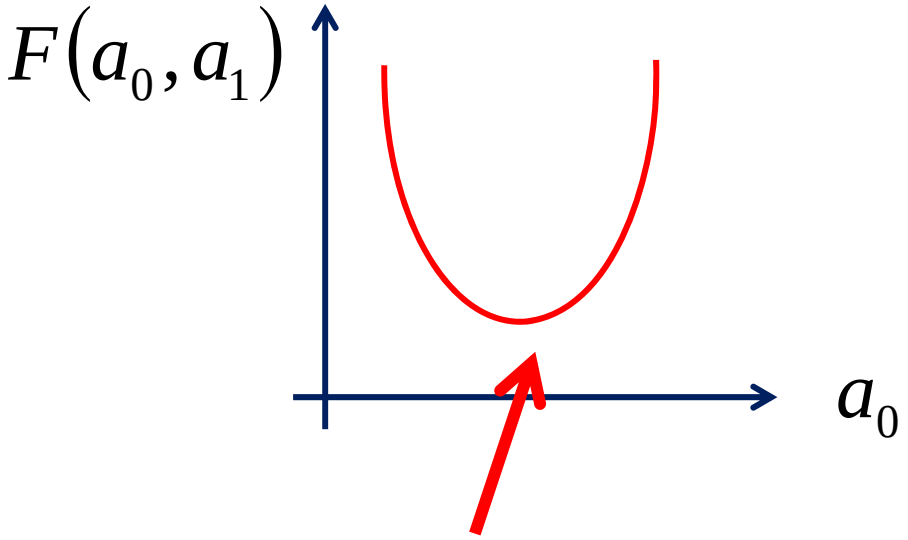
a_0, a_1 の二次関数になっている

予測誤差平方和の最小条件

- 二次関数の極値をとる条件

$$\frac{\partial}{\partial a_0} F(a_0, a_1) = 0$$

$$\frac{\partial}{\partial a_1} F(a_0, a_1) = 0$$



$$\frac{\partial}{\partial a_0} F(a_0, a_1) = 0$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_0 の条件

$$\frac{\partial}{\partial a_0} F(a_0, a_1)$$

$$= \frac{\partial}{\partial a_0} \left\{ n{a_0}^2 + {a_1}^2 \sum_{i=1}^{n} {x_i}^2 - 2a_0 \sum_{i=1}^{n} y_i - 2a_1 \sum_{i=1}^{n} y_i x_i + 2a_0 a_1 \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} {y_i}^2 \right\}$$

$$= 2na_0 - 2\sum_{i=1}^{n} y_i + 2a_1 \sum_{i=1}^{n} x_i = 0$$

$$na_0 + a_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_1 の条件

$$\frac{\partial}{\partial a_1} F(a_0, a_1)$$

$$= \frac{\partial}{\partial a_1} \left\{ n{a_0}^2 + {a_1}^2 \sum_{i=1}^{n} {x_i}^2 - 2a_0 \sum_{i=1}^{n} y_i - 2a_1 \sum_{i=1}^{n} y_i x_i + 2a_0 a_1 \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} {y_i}^2 \right\}$$

$$= 2a_1 \sum_{i=1}^{n} {x_i}^2 - 2 \sum_{i=1}^{n} y_i x_i + 2a_0 \sum_{i=1}^{n} x_i = 0$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} {x_i}^2 - \sum_{i=1}^{n} x_i y_i = 0$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_0, a_1

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n} y_i x_i = 0 \end{cases}$$

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^{\,2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_0, a_1

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_0

$$\hat{a}_0 = \frac{\sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \frac{1}{n}\sum_{i=1}^{n} x_i y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\bar{y}\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2} = \frac{\bar{y}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2\right) + \bar{x}^2\bar{y} - \bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{s_{xx}}$$

$$= \frac{\bar{y}s_{xx} - \bar{x}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}\right)}{s_{xx}} = \bar{y} - \bar{x}\frac{s_{xy}}{s_{xx}}$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_1

$$\hat{a}_1 = \frac{-\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + n\sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \overline{x}^2} = \frac{s_{xy}}{s_{xx}}$$

回帰直線

- yのxへの回帰直線

$$y = \hat{a}_0 + \hat{a}_1 x = \bar{y} - \bar{x}\frac{s_{xy}}{s_{xx}} + \frac{s_{xy}}{s_{xx}}x = \frac{s_{xy}}{s_{xx}}\left(x - \bar{x}\right) + \bar{y}$$

  - 回帰係数　$\hat{a}_1$
- xのyへの回帰直線
  - x,yの関係が逆　　$x = \frac{s_{xy}}{s_{yy}}\left(y - \bar{y}\right) + \bar{x}$

- 両者とも　$(\bar{x}, \bar{y})$ を通る。
- 傾きは異なる。　$\frac{s_{xy}}{s_{xx}}\frac{s_{xy}}{s_{yy}} \neq 1$

回帰直線の予測誤差

- yのxへの回帰直線によるyの予測値Y

$$Y = \frac{s_{xy}}{s_{xx}}\left(x - \bar{x}\right) + \bar{y}$$

  - 予測誤差 $e_i = y_i - Y_i$
  - 予測誤差の標準偏差

$$s_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^{\,2}}$$

    - 回帰式の $\hat{a}_0, \hat{a}_1$ は，目的変数を用いて求めているため，これを差し引いたn-2で割る。
    - 回帰式がk個の定数を持つ場合n-kで割る

回帰直線の予測誤差

n個のデータの
目的変数yと
説明変数xの組

	目的変数:y	説明変数:x	予測誤差
1	y_1	x_1	$e_1=y_1-(a_0+a_1x_1)$
2	y_2	x_2	$e_2=y_2-(a_0+a_1x_2)$
…			
i	y_i	x_i	$e_i=y_i-(a_0+a_1x_i)$
…			
n	y_n	x_n	$e_n=y_n-(a_0+a_1x_n)$

予測誤差の平均

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - Y_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{ y_i - \left[\frac{s_{xy}}{s_{xx}}\left(x_i - \bar{x}\right) + \bar{y} \right] \right\} = \frac{1}{n}\sum_{i=1}^{n}\left\{ y_i - \bar{y} + \frac{s_{xy}}{s_{xx}}\left(x_i - \bar{x}\right) \right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\bar{y} + \frac{s_{xy}}{s_{xx}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\bar{x} \right)$$

$$= \bar{y} - \frac{1}{n}\bar{y}n + \frac{s_{xy}}{s_{xx}}\left(\bar{x} - \frac{1}{n}\bar{x}n \right) = \bar{y} - \bar{y} + \frac{s_{xy}}{s_{xx}}\left(\bar{x} - \bar{x}\right) = 0$$

相関係数

- 変量間の関係の強さを表す
  - 変量x,y間の相関係数 $r_{xy} = \dfrac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$
    - 決定係数 $r_{xy}^{\ 2}$
- 予測誤差の標準偏差と相関係数の関係

$$s_e^{\ 2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^{\ 2} = \frac{n}{n-2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - Y_i\right)^2$$

$$= \frac{n}{n-2}\frac{1}{n}\sum_{i=1}^{n}\left\{ y_i - \left[\frac{s_{xy}}{s_{xx}}\left(x_i - \bar{x}\right) + \bar{y} \right] \right\}^2$$

相関係数

$$s_e^{\,2} = \frac{n}{n-2}\frac{1}{n}\sum_{i=1}^{n}\left\{ y_i - \bar{y} - \frac{s_{xy}}{s_{xx}}\left(x_i - \bar{x}\right)\right\}^2$$

$$= \frac{n}{n-2}\frac{1}{n}\left\{ \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 - 2\frac{s_{xy}}{s_{xx}}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) + \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \right\}$$

$$= \frac{n}{n-2}\left\{ s_{yy} - 2\frac{s_{xy}}{s_{xx}}s_{xy} + \left(\frac{s_{xy}}{s_{xx}}\right)^2 s_{xx} \right\} = \frac{n}{n-2}\left\{ s_{yy} - \frac{s_{xy}^{\,2}}{s_{xx}} \right\}$$

$$= \frac{n}{n-2}s_{yy}\left\{ 1 - \frac{s_{xy}^{\,2}}{s_{xx}s_{yy}} \right\} = \frac{n}{n-2}s_{yy}\left\{ 1 - r_{xy}^{\,2} \right\}$$