

応用システム工学

第四回 回帰分析

平成26年05月09日

直線回帰

線形重回帰

説明変数の種類

- 一般化線形モデル

- 反応変数 Y と説明変数 x_1, x_2, \dots, x_m の関係

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$$

----- 線形関数

- 量的な説明変数 x

- パラメータ β は, 説明変数の変化に伴う反応変数の変化を表す

- 質的な説明変数

- 反応変数にパラメータが含まれるか否かを表す
 - ダミー変数, $(0, 1)$ の場合は指示変数

多変量解析 ⇒ 相関分析, 回帰分析

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

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

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

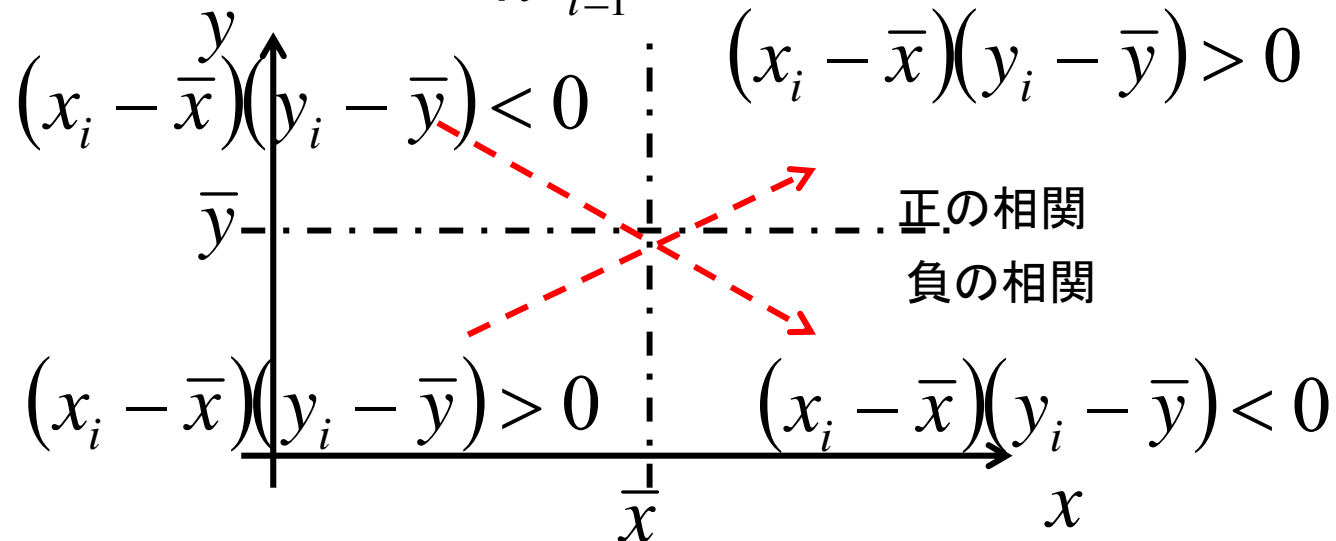
$$\rho_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad \text{は各々の相加平均}$$
$$|\rho_{xy}| < 1$$

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

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



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
(ある変数の変化をもう一方の変数の変化で説明するための関数を求める)

$$a_0 + a_1x_1 + \cdots + a_px_p \rightarrow y$$

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

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

- 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

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

- 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
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- 目的変数
$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

線形回帰モデルのパラメータ同定

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

• 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

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

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)

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

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

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

• 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

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

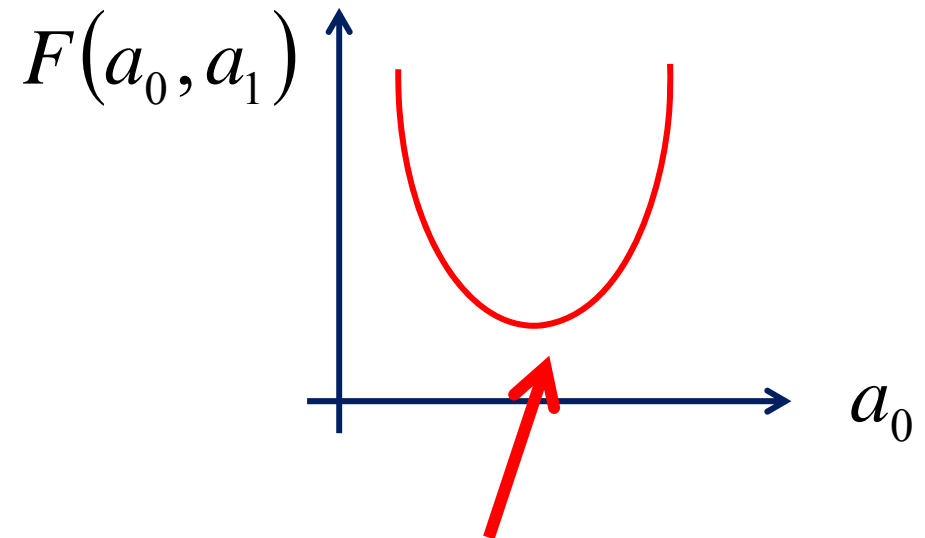
a_0, a_1 の二次関数になっている

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

$$\begin{aligned} & \frac{\partial}{\partial a_0} F(a_0, a_1) \\ &= \frac{\partial}{\partial a_0} \left\{ na_0^2 + a_1^2 \sum_{i=1}^n x_i^2 - 2a_0 \sum_{i=1}^n y_i - 2a_1 \sum_{i=1}^n y_i x_i + 2a_0 a_1 \sum_{i=1}^n x_i + \sum_{i=1}^n y_i^2 \right\} \\ &= 2na_0 - 2 \sum_{i=1}^n y_i + 2a_1 \sum_{i=1}^n x_i = 0 \\ & \quad na_0 + a_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0 \end{aligned}$$

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

$$\begin{aligned} & \frac{\partial}{\partial a_1} F(a_0, a_1) \\ &= \frac{\partial}{\partial a_1} \left\{ na_0^2 + a_1^2 \sum_{i=1}^n x_i^2 - 2a_0 \sum_{i=1}^n y_i - 2a_1 \sum_{i=1}^n y_i x_i + 2a_0 a_1 \sum_{i=1}^n x_i + \sum_{i=1}^n y_i^2 \right\} \\ &= 2a_1 \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n y_i x_i + 2a_0 \sum_{i=1}^n x_i = 0 \end{aligned}$$

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

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

$$\begin{aligned} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} &= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & - \sum_{i=1}^n x_i \\ - \sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \end{aligned}$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_0

$$\begin{aligned}
 \hat{a}_0 &= \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\frac{1}{n} \sum_{i=1}^n x_i^2 \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n x_i \frac{1}{n} \sum_{i=1}^n x_i y_i}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2} \\
 &= \frac{\bar{y} \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x} \frac{1}{n} \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\bar{y} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right) + \bar{x}^2 \bar{y} - \bar{x} \frac{1}{n} \sum_{i=1}^n x_i y_i}{S_{xx}} \\
 &= \frac{\bar{y} S_{xx} - \bar{x} \left(\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \right)}{S_{xx}} = \bar{y} - \bar{x} \frac{S_{xy}}{S_{xx}}
 \end{aligned}$$

予測誤差平方和の最小条件

- 二次関数の極値をとる a_1

$$\begin{aligned}\hat{a}_1 &= \frac{-\sum_{i=1}^n x_i \sum_{i=1}^n y_i + n \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \frac{1}{n} \sum_{i=1}^n y_i}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{s_{xy}}{s_{xx}}\end{aligned}$$

回帰直線

- y の x への回帰直線

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

– 回帰係数 \hat{a}_1

- x の y への回帰直線

– x, y の関係が逆 $x = \frac{S_{xy}}{S_{yy}} (y - \bar{y}) + \bar{x}$

- 両者とも (\bar{x}, \bar{y}) を通る。

- 傾きは異なる。 $\frac{S_{xy}}{S_{xx}} \frac{S_{xy}}{S_{yy}} \neq 1$