

応用システム工学 第一回 回帰分析

平成23年04月08日

2011/04/08

1

変数

- 原因
 - 説明変数
 - 予測変数
 - 独立変数
- 結果
 - 目的変数
 - 反応変数
 - 結果変数
 - 従属変数

2011/04/08

2

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

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

2011/04/08

3

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

2011/04/08

4

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

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

線形回帰モデルのパラメータ同定

$$\begin{aligned} \text{— 分散 } s_{xx} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \bar{x}^2 \sum_{i=1}^n 1 \right) \\ &= \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 \right) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \end{aligned}$$

• 同様に

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

2011/04/08

7

線形回帰モデルのパラメータ同定

$$\begin{aligned} \text{— 共分散} \\ s_{xy} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\ &= \frac{1}{n} \left(\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + \bar{x} \bar{y} \sum_{i=1}^n 1 \right) \\ &= \frac{1}{n} \left(\sum_{i=1}^n x_i y_i - \bar{y}n\bar{x} - \bar{x}n\bar{y} + \bar{x}\bar{y}n \right) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y} \end{aligned}$$

• 同様に $s_{yx} = s_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$

2011/04/08

8

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

2011/04/08

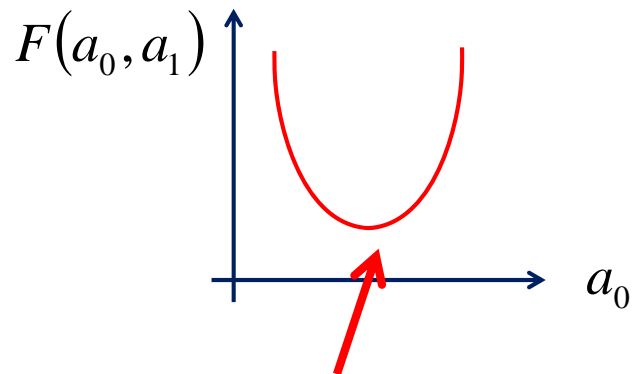
9

予測誤差平方和の最小条件

- 二次関数の極値をとる条件

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

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



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

2011/04/08

10

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

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

予測誤差平方和の最小条件

- 二次関数の極値をとる a_0, 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

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

2011/04/08 15

予測誤差平方和の最小条件

- 二次関数の極値をとる a_1

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

2011/04/08 16

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, y の関係を逆にした x の y への回帰直線

$$x = \frac{s_{xy}}{s_{yy}} (y - \bar{y}) + \bar{x}$$

- 両者とも (\bar{x}, \bar{y}) を通る。

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

2011/04/08

17

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

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

– 予測誤差 $e_i = y_i - Y_i$

– 予測誤差の標準偏差

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

- 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 $n-2$ で割っている。
- 回帰式が k 個の定数を持つ場合 $n-k$ で割る

2011/04/08

18

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	$e_n=y_n-(a_0+a_1x_n)$

2011/04/08

19

予測誤差の平均

$$\begin{aligned}\bar{e} &= \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n (y_i - Y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ y_i - \left[\frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) + \bar{y} \right] \right\} = \frac{1}{n} \sum_{i=1}^n \left\{ y_i - \bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right\} \\ &= \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n \bar{y} + \frac{s_{xy}}{s_{xx}} \left(\frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n \bar{x} \right) \\ &= \bar{y} - \frac{1}{n} \bar{y}n + \frac{s_{xy}}{s_{xx}} \left(\bar{x} - \frac{1}{n} \bar{x}n \right) = \bar{y} - \bar{y} + \frac{s_{xy}}{s_{xx}} (\bar{x} - \bar{x}) = 0\end{aligned}$$

2011/04/08

20

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 $r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$

- 予測誤差の標準偏差と相関係数の関係

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

相関係数

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