回帰分析 (regression analysis)

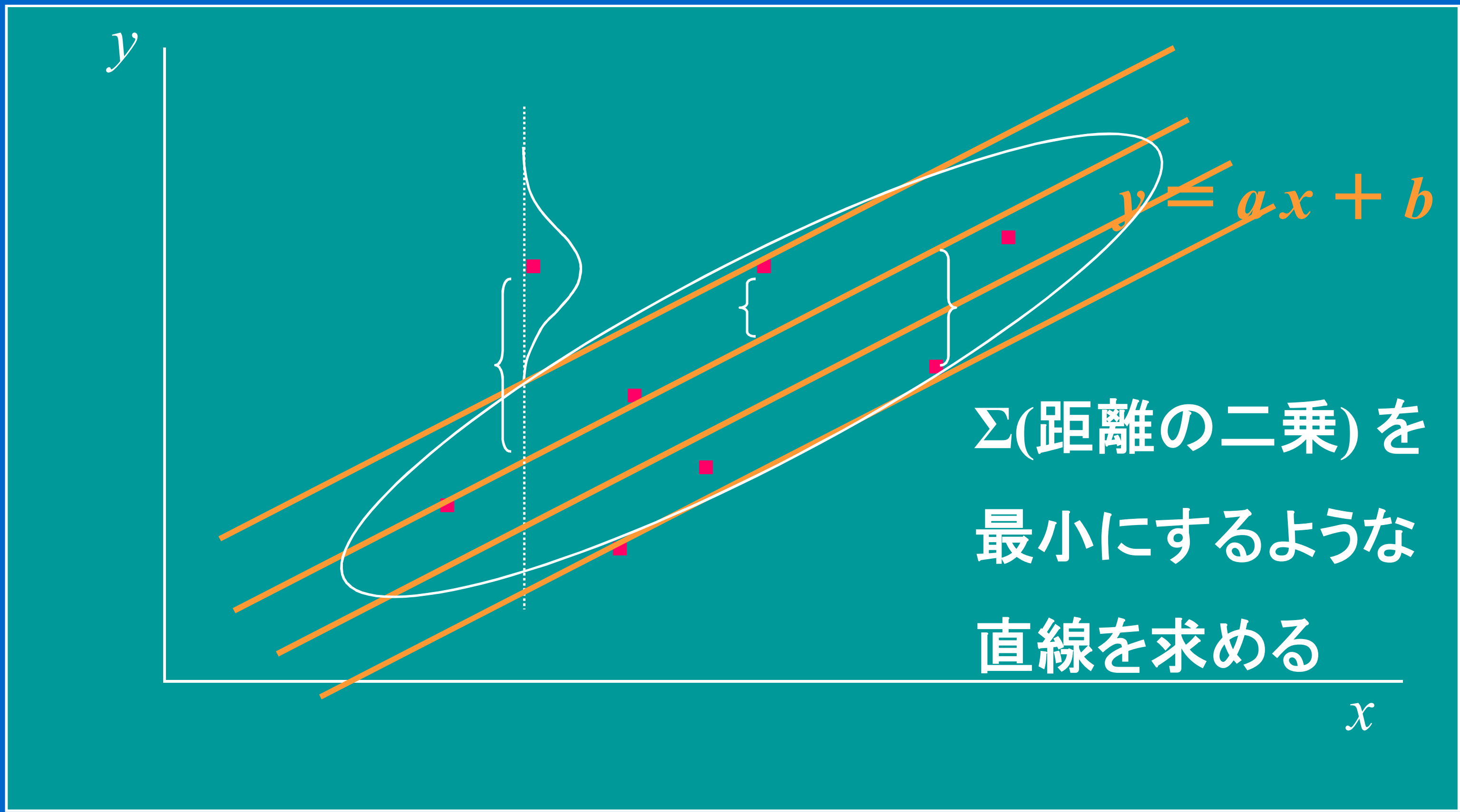
n組の原因となる数値(複数個でもよい)と結果の数値から，両者の関係を示す直線あるいは曲線を求める.

独立変数(説明変数) ・・・・・ 原因となる数値
従属変数(被説明変数，応答変数，目的変数) ・・・ 結果となる数値

単回帰 ・・・ 独立変数が1つ $y = a x + b$
重回帰 ・・・ 独立変数が2つ以上 $y = a x_1 + b x_2 + c$

回帰直線 ・・・ 回帰分析により求める直線 $y = a x + b$
回帰曲線 ・・・ 回帰分析により求める曲線 $y = a x^2 + b x + c$

最小二乗法 (method of least squares)



最小二乗法による直線回帰

従属変数(y)	独立変数(x)	(x^2)	($x\,y$)
y_1 ・ y_i ・ y_n	x_1 ・ x_i ・ x_n	$x_1^{\,2}$ ・ $x_i^{\,2}$ ・ $x_n^{\,2}$	$x_1\,y_1$ ・ $x_i\,y_i$ ・ $x_n\,y_n$
Σy_i	Σx_i	$\Sigma x_i^{\,2}$	$\Sigma x_i\,y_i$

から, $y = a\,x + b$ の a , b を求める.

$$\begin{cases} a\Sigma x_i + nb = \Sigma y_i \\ a\Sigma x_i^{\,2} + b\Sigma x_i = \Sigma x_i\,y_i \end{cases}$$

（正規方程式）

$$a = \frac{\Sigma x_i\,y_i - \dfrac{\Sigma x_i\,\Sigma y_i}{n}}{\Sigma x_i^{\,2} - \dfrac{(\Sigma x_i)^2}{n}}$$

$$b = \overline{y} - a\,\overline{x}$$

a を求めるための別法

従属変数(y)	独立変数(x)	$(x_i-\overline{x})$	$(y_i-\overline{y})$	$(x_i-\overline{x})(y_i-\overline{y})$	$(x_i-\overline{x})^2$
y_1 . y_i . y_n	x_1 . x_i . x_n	$(x_1-\overline{x})$. $(x_i-\overline{x})$. $(x_n-\overline{x})$	$(y_1-\overline{y})$. $(y_i-\overline{y})$. $(y_n-\overline{y})$	$(x_1-\overline{x})(y_1-\overline{y})$. $(x_i-\overline{x})(y_i-\overline{y})$. $(x_n-\overline{x})(y_n-\overline{y})$	$(x_1-\overline{x})^2$. $(x_i-\overline{x})^2$. $(x_n-\overline{x})^2$
Σy_i	Σx_i			Σ Σ	

から, $y=ax+b$ の a , b を求める.

$$\begin{cases} a\Sigma x_i + b = \Sigma y_i \\ a\Sigma x_i^2 + b\Sigma x_i = \Sigma x_i y_i \end{cases}$$

（正規方程式）

$$a=\frac{\Sigma(x_i-\overline{x})(y_i-\overline{y})}{\Sigma(x_i-\overline{x})^2}$$

$$b=\overline{y}-a\overline{x}$$

例題

西暦	平均気温(x)	収穫量(y)	(x^2)	($x\,y$)
1990	25	30	625	750
1991	29	35	841	1015
1992	35	40	1225	1400
1993	20	28	400	560
1994	36	42	1296	1512
合計	145 Σx	175 Σy	4387 Σx^2	5237 Σxy
平均	29	35		

求める回帰直線は

$$y = 0.89\,x + 9.19$$ //

平均気温が27度のときの予想収穫量は,

$x = 27$ より

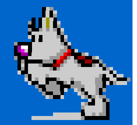
$y = 33.13$ //

$$a = \frac{5237 - \dfrac{145 \times 175}{5}}{4387 - \dfrac{145^2}{5}} \fallingdotseq 0.89$$

$$b = 35 - 0.89 \times 29$$
$$= 9.19$$

分析ツール

	A	B	C	D	E	F	G	H	I
4	重相関 R	0.987071							
5	重決定 R2	0.974309							
6	補正 R2	0.965746							
7	標準誤差	1.125788							
8	観測数	5							
9									
10	分散分析表								
11		自由度	変動	分散	訓された分散	有意 F			
12	回帰	1	144.1978	144.1978	113.7746	0.001761			
13	残差	3	3.802198	1.267399					
14	合計	4	148						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	9.186813	2.471837	3.716593	0.033887	1.320324	17.0533	1.320324	17.0533
18	X 値 1	0.89011	0.083449	10.66652	0.001761	0.624538	1.155682	0.624538	1.155682
19									



例題 - 2

学生	英語(x)	数学(y)	(x^2)	$(x\,y)$
1	40	50	1600	2000
2	40	30	1600	1200
3	60	50	3600	3000
4	60	70	3600	4200
5	80	70	6400	5600
6	80	90	6400	7200
合計	360	360	23200	23200

$$a = \frac{23200 - \dfrac{360 \times 360}{6}}{23200 - \dfrac{360^{2}}{6}}$$

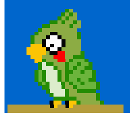
$$= 1$$

$$b = 60 - 1 \times 60$$

$$= 0$$

求める回帰直線は

$$y = x$$



回帰係数の検定（回帰式の有効性）

$\hat{y}_i$ は回帰式 $\hat{y}=ax+b$ の x に x_i を代入したときの
y の予測値 ax_i+b を表す.

手順

①自由度 ($\varphi=n-2$) を求める

② $t_0= a\sqrt{\dfrac{\varphi\Sigma(x_i-x)^2}{\Sigma(y_i-\hat{y})^2}}$

③ t 分布表より，自由度φ，有意水準$\alpha/2$ の点$t(\varphi,\alpha/2)$を知る

④ $|t_0| \geqq t(\varphi/2)$ のとき，

母回帰係数 $a=0$ という帰無仮説は棄却される.

すなわち，求めた回帰式は予測に役立つ.

$|t_0| < t(\varphi/2)$ のとき，帰無仮説は棄却されない.

すなわち，求めた回帰式は予測に役立たない.

例題

平均気温と収穫量の問題で

$a = 0.89$ ， $b = 9.19$ であった.

回帰係数の検定を5％の有意水準で行え.

$\hat{y}_i = 0.89x_i + 9.19$ (回帰直線)

例題の解答例

	平均気温(x)	収穫量(y)	$(x_i - \bar{x})^2$	($\hat{y}$)	$(y_i - \hat{y}_i)^2$
	25	30	16	31.44	2.0736
	29	35	0	35.00	0
	35	40	36	40.34	0.1156
	20	28	81	26.99	1.0201
	36	42	49	41.23	0.5929
平均	29	35			
合計			182	175	3.8022

自由度φ

$= 5 - 2 = 3$

$\Sigma(x_i - \bar{x})^2$

$= 182$

$\Sigma(y_i - \hat{y}_i)^2$

$= 3.8022$

$$t_0 = 0.89 \times \sqrt{\frac{3 \times 182}{3.8022}}$$

$\fallingdotseq 10.67$

$t\,(3,\ 0.025) = 3.182$　　$t_0 = 10.67 > 3.182$

∴ 回帰式は予測に役立つ

すなわち，平均気温から収穫量の予測が可能

決定係数

回帰直線のあてはまりの良さを表す指標

$$R^2 = \frac{\text{回帰変動}}{\text{全変動}}$$

$$0 \leqq R^2 \leqq 1$$

【例題】 平均気温と収穫量の場合

$R^2 = 144.1622 / 148$

$\fallingdotseq 0.974$

$\therefore$ 回帰直線によるあてはまりが極めて良い //

分析ツール

	A	B	C	D	E	F	G	H	I
4	重相関 R	0.987071							
5	重決定 R2	0.974309							
6	補正 R2	0.965746							
7	標準誤差	1.125788							
8	観測数	5							
9									
10	分散分析表								
11		自由度	変動	分散	された分散	有意 F			
12	回帰	1	144.1978	144.1978	113.7746	0.001761			
13	残差	3	3.802198	1.267399					
14	合計	4	148						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	9.186813	2.471837	3.716593	0.033887	1.320324	17.0533	1.320324	17.0533
18	X 値 1	0.89011	0.083449	10.66652	0.001761	0.624538	1.155682	0.624538	1.155682
19									