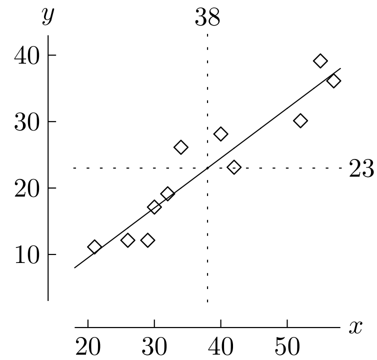
6　回帰直線

6.1　回帰直線

次のデータは種目 X の記録 x（kgw）と種目 Y の記録 y（m）である。

x	21	26	29	30	32	34	40	42	52	55	57
y	11	12	12	17	19	26	28	23	30	39	36

散布図上で，誤差が最小となるように引いた直線のことを**回帰直線**という。



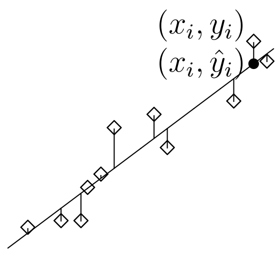
回帰直線の方程式を**回帰式**（y の x への回帰式）という。

$$\hat{y} = a + bx \tag{1}$$

回帰式 (1) において，a を回帰定数，b を回帰係数という。また x を説明変数，y を被説明変数または目的変数という。説明変数の値 x_i に対して回帰式 (1) から求めた値 $\hat{y}_i = a + bx_i$ のことを**回帰値**という。

最小二乗法（least squares method）　観測値 y_i と回帰値 $\hat{y}_i$ の差を**残差**または誤差という。残差を e_i とおく。

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i$$



残差平方和を $S(a,b)$ とおく [*i]。最小二乗法とは，残差平方和が最小となるように定数 a，b を定める手法のことである。　*i

$$S(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

[*i] 残差 e_i の 2 乗の総和を残差平方和（RSS; Residual sum of squares）という。

6.2 回帰式の導出

残差平方和 $S(a,b)$ を a，b について整理し，最小二乗解を求める。

$$\begin{aligned} S(a,b) &= \sum (y_i - a - bx_i)^2 \\ &= \sum \left\{ \bar{y} - a - b\bar{x} + (y_i - \bar{y}) - b(x_i - \bar{x}) \right\}^2 \\ &= \sum (\bar{y} - a - b\bar{x})^2 + \sum \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\}^2 \\ &\qquad + 2(\bar{y} - a - b\bar{x}) \sum \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\} \qquad (*) \end{aligned}$$

*ii 偏差の総和は 0 なので [*ii]，$(*)$ の項は 0 になる。

$$\begin{aligned} S(a,b) &= \sum (\bar{y} - a - b\bar{x})^2 + \sum \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\}^2 \\ &= \sum (\bar{y} - a - b\bar{x})^2 + b^2 \sum (x_i - \bar{x})^2 \\ &\qquad - 2b \sum (x_i - \bar{x})(y_i - \bar{y}) + \sum (y_i - \bar{y})^2 \\ &= n(\bar{y} - a - b\bar{x})^2 + b^2 n s_x^2 - 2bn s_{xy} + n s_y^2 \end{aligned}$$

*iii 残差平方和 $S(a,b)$ は次のように変形できる。ただし $s_x^2 > 0$ とする。[*iii]

$$S(a,b) = n(\bar{y} - a - b\bar{x})^2 + n s_x^2 \left(b - \frac{s_{xy}}{s_x^2} \right)^2 + n \frac{s_x^2 s_y^2 - {s_{xy}}^2}{s_x^2}$$

一般に，平方数は負になることはない（$X^2 \geqq 0$）から，残差平方和 $S(a,b)$ の最小値は $n(s_x^2 s_y^2 - {s_{xy}}^2)/s_x^2$ で，それが最小となるのは $\bar{y} - a - b\bar{x} = 0$ かつ $b - s_{xy}/s_x^2 = 0$ のときである。回帰式 (1) において a，b は次の値となる。

$$b = \frac{s_{xy}}{s_x^2}, \qquad a = \bar{y} - b\bar{x} \tag{2}$$

回帰直線と点 $(\bar{x}, \bar{y})$　回帰式 (1)，(2) から，$\hat{y} = (\bar{y} - b\bar{x}) + bx$ となるから，

$$\hat{y} - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \tag{3}$$

と表せる。この (3) 式の形から，回帰直線は必ず点 $(\bar{x}, \bar{y})$ を通る。

[*ii] $\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$

[*iii] $s_x^2 = 0$ のときは，最小二乗解 a，b が定まらない。

6.3 回帰直線の求め方

回帰直線の求め方（1）　上のデータに対して回帰式 $\hat{y} = a + bx$ と，すべての回帰値 $\hat{y}_i$ を求める。

変数	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	$\hat{y}$
a.	**21**	**11**	-17	-12	289	144	204	**10.25**
b.	**26**	**12**	-12	-11	144	121	132	**14.00**
c.	**29**	**12**	-9	-11	81	121	99	**16.25**
d.	**30**	**17**	-8	-6	64	36	48	**17.00**
e.	**32**	**19**	-6	-4	36	16	24	**18.50**
f.	**34**	**26**	-4	3	16	9	-12	**20.00**
g.	**40**	**28**	2	5	4	25	10	**24.50**
h.	**42**	**23**	4	0	16	0	0	**26.00**
i.	**52**	**30**	14	7	196	49	98	**33.50**
j.	**55**	**39**	17	16	289	256	272	**35.75**
k.	**57**	**36**	19	13	361	169	247	**37.25**
合計	418	253	0	0	1496	946	1122	253
平均	38	23	0	0	136	86	102	23

平均，分散，共分散は，

$$\bar{x} = 38, \qquad \bar{y} = 23, \qquad s_x^2 = 136, \qquad s_{xy} = 102$$

回帰係数 b，回帰定数 a は，

$$b = \frac{s_{xy}}{s_x^2} = \frac{102}{136} = 0.75, \qquad a = \bar{y} - b\bar{x} = 23 - \frac{102}{136} \cdot 38 = -5.5$$

なので，回帰式 $\hat{y} = a + bx$ は次のとおり。または $\hat{y} - \bar{y} = b(x - \bar{x})$ でもよい。

$$\hat{y} = -5.5 + 0.75x, \qquad \hat{y} - 23 = 0.75(x - 38)$$

$x_1 = 21$ に対応する回帰値は $\hat{y}_1 = 10.25$ である（他の回帰値も同様にする）。

$$\hat{y}_1 = -5.5 + 0.75 \cdot 21 = 10.25$$

回帰値の平均値　回帰式 (1)，(2) と，平均値の性質から，

$$\bar{\hat{y}} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}$$

が成り立つので，回帰値の平均値 $\bar{\hat{y}}$ は観測値の平均値 $\bar{y}$ に等しい。

$$\bar{\hat{y}} = \bar{y} \tag{4}$$

回帰直線の求め方（2）　回帰式 $\hat{y} = a + bx$ は次のように求めることもできる。

変数	x	y	x^2	y^2	xy	$\hat{y}$
a.	**21**	**11**	441	121	231	**10.25**
b.	**26**	**12**	676	144	312	**14.00**
c.	**29**	**12**	841	144	348	**16.25**
d.	**30**	**17**	900	289	510	**17.00**
e.	**32**	**19**	1024	361	608	**18.50**
f.	**34**	**26**	1156	676	884	**20.00**
g.	**40**	**28**	1600	784	1120	**24.50**
h.	**42**	**23**	1764	529	966	**26.00**
i.	**52**	**30**	2704	900	1560	**33.50**
j.	**55**	**39**	3025	1521	2145	**35.75**
k.	**57**	**36**	3249	1296	2052	**37.25**
合計	418	253	17380	6765	10736	253
平均	38	23	1580	615	976	23

平均は $\bar{x} = 66$，$\bar{y} = 69$ である。分散，共分散は，

$$s_x^2 = 1580 - 38^2 = 136, \qquad s_{xy} = 976 - 38 \cdot 23 = 102$$

回帰係数 b，回帰定数 a は，

$$b = \frac{s_{xy}}{s_x^2} = \frac{102}{136} = 0.75, \qquad a = \bar{y} - b\bar{x} = 23 - \frac{102}{136} \cdot 38 = -5.5$$

なので，回帰式 $\hat{y} = a + bx$ は次のとおり。または $\hat{y} - \bar{y} = b(x - \bar{x})$ でもよい。

$$\hat{y} = -5.5 + 0.75x, \qquad \hat{y} - 23 = 0.75(x - 38)$$

回帰直線の求め方（3）　「回帰直線の求め方（2）」の表を利用すると，平均は $\bar{x} = 418/11 = 38$，$\bar{y} = 253/11 = 23$ である。分散，共分散は，

$$s_x^2 = \frac{11 \cdot 17380 - 418^2}{11^2} = \frac{16456}{11^2}, \quad s_{xy} = \frac{11 \cdot 10736 - 418 \cdot 253}{11^2} = \frac{12342}{11^2}$$

回帰係数 b，回帰定数 a は，

$$b = \frac{n^2 \cdot s_{xy}}{n^2 \cdot s_x^2} = \frac{12342}{16456} = 0.75, \qquad a = \bar{y} - b\bar{x} = 23 - \frac{12342}{16456} \cdot 38 = -5.5$$

なので，回帰式 $\hat{y} = a + bx$ は次のとおり。または $\hat{y} - \bar{y} = b(x - \bar{x})$ でもよい。

$$\hat{y} = -5.5 + 0.75x, \qquad \hat{y} - 23 = 0.75(x - 38)$$

6.4　微分係数と極値

1 変数関数の極小点　関数 $y = f(x)$ は微分可能で，導関数 $f'(x)$ は連続とする。

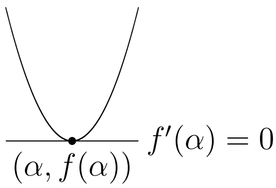
$$f'(x) = \frac{d}{dx}f(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

導関数を求めるための幾つかの公式が知られている。

$$(x^2)' = 2x, \qquad \{(g(x))^2\}' = 2(g(x)) \times g'(x)$$

極小点とは，その近傍において最小となる点のことである。微分係数 $f'(\alpha)$ は点 $(\alpha, f(\alpha))$ における接線の傾きを表し，極小点に関して次の命題が成り立つ。

$$f(x) \text{ が } x = \alpha \text{ で極小} \implies f'(\alpha) = 0$$



2 変数関数の極小点　2 変数関数 $z = f(x, y)$ は偏微分可能で，偏導関数 $f_x(x, y)$，$f_y(x, y)$ は連続とする。

$$f_x(x,y) = \frac{\partial}{\partial x}f(x,y) = \lim_{h\to 0}\frac{f(x+h,y)-f(x,y)}{h}$$
$$f_y(x,y) = \frac{\partial}{\partial y}f(x,y) = \lim_{k\to 0}\frac{f(x,y+k)-f(x,y)}{k}$$

次の命題が成り立つ。

$$f(x,y) \text{ が } x = \alpha,\ y = \beta \text{ で極小} \implies f_x(\alpha,\beta) = f_y(\alpha,\beta) = 0$$

6.5　偏微分法による回帰式の導出

被説明変数を 1 個の説明変数で表した式のことを単回帰式という [*iv]。単回帰式は回帰直線を作る。単回帰式を $\hat{y} = a + bx$ とし，残差平方和を $S(a, b)$ とおく。　*iv

$$S(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

[*iv] 被説明変数を 2 個以上の説明変数で表した式のことを重回帰式という。

正規方程式　残差平方和 $S(a,b)$ を偏微分し，最小二乗解を求める。

$$S_a(a,b) = -2\sum(y_i - a - bx_i) = 0$$
$$S_b(a,b) = -2\sum x_i(y_i - a - bx_i) = 0$$

回帰定数 a，回帰係数 b は次の連立方程式を満たす。これを正規方程式という。

$$\sum(y_i - a - bx_i) = 0$$
$$\sum x_i(y_i - a - bx_i) = 0$$

単回帰式の導出　正規方程式をデータの大きさ n で割り，平均値の式で表す。

$$\bar{y} - a - b\bar{x} = 0 \tag{i}$$
$$\overline{xy} - a\bar{x} - b\overline{x^2} = 0 \tag{ii}$$

(ii) から (i) $\times\ \bar{x}$ を引き，文字 a を消去する。

$$\overline{xy} - \bar{x}\bar{y} - b(\overline{x^2} - \bar{x}^2) = 0$$

$s_{xy} = \overline{xy} - \bar{x}\bar{y}$，$s_x^2 = \overline{x^2} - \bar{x}^2$ を用いると，次の公式が得られる。

$$\hat{y} = a + bx, \qquad b = \frac{s_{xy}}{s_x^2}, \qquad a = \bar{y} - b\bar{x} \tag{5}$$

本稿の参考文献

統計学入門（基礎統計学）
東京大学教養学部統計学教室（編）東京大学出版会 978-4-13-042065-5

統計学
久保川 達也 （著） 東京大学出版会 978-4-13-062921-8

はじめての統計学
道家 暎幸（共著） コロナ社 978-4-339-06113-0

確率統計 新版 （新版数学シリーズ）
岡本 和夫 （ほか著） 実教出版 978-4-407-32171-5

統計学序論 改訂版
山本 義郎 （著） 東海大学出版部 978-4-486-02133-9

確率統計（高専テキストシリーズ）
上野 健爾（監修） 森北出版 978-4-627-05561-2

基本統計学 第４版
宮川 公男 （著） 有斐閣 978-4-641-16455-0

新統計入門
小寺 平治 （著） 裳華房 978-4-7853-1099-8

Schaum's Outline of Introduction to Probability and Statistics
Seymour Lipschutz （著） McGraw-Hill Education 978-0-07-176249-6

A Dictionary of Statistics
Graham Upton （著） Oxford Univ Pr 978-0-19-967918-8

www5e.biglobe.ne.jp/~emm386/statistics/