● 回帰分析

2 つの変量 X と Y の間に相関関係があるか，またその強さがどれくらいかを測るのが相関係数であった．少し観点を変え，相関関係があると思われる 2 つの変量のうち、一方の変量 X から他方の変量 Y の値がどの程度推測できるかを考えるのが回帰分析である．

まず，2 つの変量 X，Y をそれぞれ確率変数と捉え，X を説明変数，Y を目的変数と呼ぶ．回帰分析とは，説明変数と目的変数の関係を回帰式で表し、目的変数が説明変数によってどの程度説明できるかを定量的に分析することである．回帰分析の最も基本的なモデルは回帰式が $Y = a + bX$ という形式の一次式で表せる線形回帰である。

例．ある大学の定期試験の結果から 10 人の学生の成績を無作為に抽出してみたところ，数学と物理の点は下の表のようであった．

学生	1	2	3	4	5	6	7	8	9	10
数学 : X(点)	79	63	78	86	65	58	93	83	75	70
物理 : Y(点)	85	70	82	83	75	70	90	77	76	78

まず，仮平均を数学，物理ともに 75 点として，$U = X - 75$，$V = Y - 75$ とおく．すでに示したように，

(1)
$$E(X) = E(U) - 75, \qquad E(Y) = E(V) - 75,$$
$$V(X) = V(U) = E(U^2) - E(U)^2, \quad V(Y) = V(V) = E(V^2) - E(V)^2$$

が成り立ち，X と Y の共分散 $\mathrm{Cov}(X, Y)$ についても同様に

(2)
$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(U, V) = E(UV) - E(U)E(V)$$

が成り立つ．これらを計算するために，次のような表をまず作る．

	X	Y	U	V	U^2	V^2	UV
1	79	85	4	10	16	100	40
2	63	70	−12	−5	144	25	60
3	78	82	3	7	9	49	63
4	86	83	11	8	121	64	88
5	65	75	−10	0	100	0	0
6	58	70	−17	−5	289	25	85
7	93	90	18	15	324	225	270
8	83	77	8	2	64	4	16
9	75	76	0	1	0	1	0
10	70	78	−5	3	25	9	-15
		和	0	36	1092	502	565
		平均	0	3.6	109.2	50.2	56.5

この結果から，(1)，(2) を用いて

$$V(X) = 109.2, \quad V(Y) = 37.24, \quad \mathrm{Cov}(X, Y) = 56.5$$

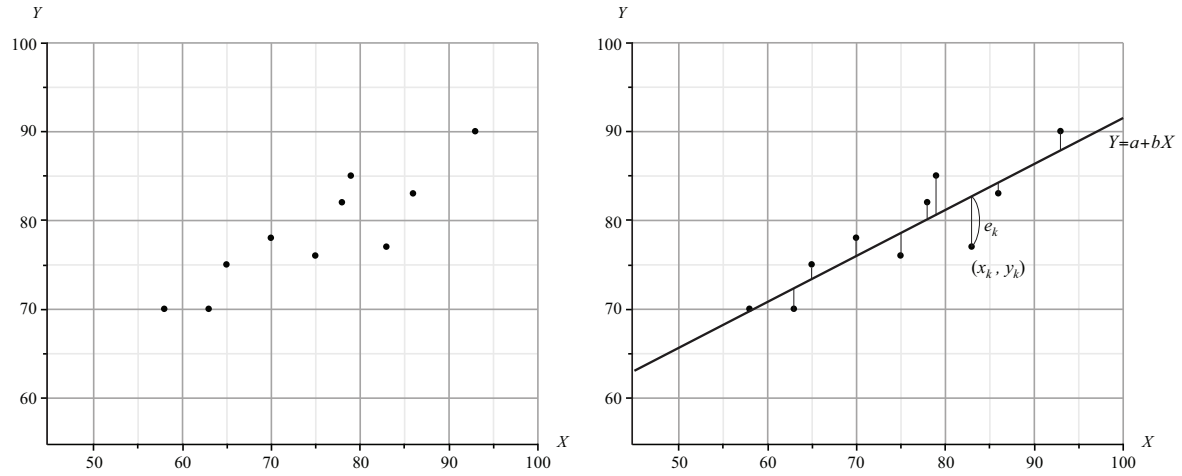
となる．これより，X と Y の間の相関係数 r は

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma(X)\sigma(Y)} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{56.5}{\sqrt{109.2}\sqrt{37.24}} \fallingdotseq 0.886$$

となり，比較的強い正の相関関係があることがわかる．

【裏につづく】

次に，X と Y の 10 個の標本の値 $(x_1, y_1), (x_2, y_2), \ldots (x_{10}, y_{10})$ をグラフ上の点として表すと，下の左の図のような散布図が得られる．点の全体は右上がりの向きに分布している．この場合に，できるだけ客観的に直線を当てはめたい．その時に用いられるのが最小 2 乗法である．



いま，X と Y の間に近似的に $Y = a + bX$ という直線関係が成り立っているとする．（$Y = aX + b$ ではないことに注意.）この関係は近似的な関係なので，$X = x_k$ のときに $Y = y_k$ と $a + bx_k$ との間に誤差

$$e_k = y_k - (a + bx_k)$$

が生じる．これらの誤差をできるだけ小さくするようにするにはいろいろな方法が考えられるが，e_k^2 の平均

$$Q = \frac{1}{10}(e_1^2 + e_2^2 + \cdots + e_{10}^2) = \frac{1}{10}\sum_{k=1}^{10}\big(y_k - (a + bx_k)\big)^2$$

を最小にするように a，b を定めるのが最小 2 乗法である．この方法によって得られる結果を先に述べると，

$$a = E(Y) - \frac{\mathrm{Cov}(X,Y)}{V(X)}E(X), \quad b = \frac{\mathrm{Cov}(X,Y)}{V(X)} \tag{3}$$

となることが知られている．これは，$Y = a + bX$ が点 $\big(E(X), E(Y)\big)$ を通り，傾きが上の b の値であるような直線であることを意味している．具体的な数値を計算してみると

$$a = 78.6 - \frac{56.5}{109.2} \times 75.0 \fallingdotseq 39.8, \quad b = \frac{56.5}{109.2} \fallingdotseq 0.52$$

となる．これにより，Y の値を X の値により推定する式 $Y = 39.8 + 0.52X$ が得られる．これより，数学の点数が例えば 80 点だった学生の物理の点数はだいたい 81.4 点だったであろうと推測できる．

参考．少々技巧的だが，(3) は次のように得られる．最小にすべき Q は $E\Big(\big(Y - (a + bX)\big)^2\Big)$ と書ける．そこで，$E(X) = \bar{x}$，$E(Y) = \bar{y}$ とおいて，$Y - (a + bX) = (Y - \bar{y}) - b(X - \bar{x}) + (\bar{y} - a - b\bar{x})$ と書き直すと

$$\begin{aligned}
&E\Big(\big(Y - (a + bX)\big)^2\Big) = E\Big(\big((Y - \bar{y}) - b(X - \bar{x}) + (\bar{y} - a - b\bar{x})\big)^2\Big) \\
&= E\Big((Y - \bar{y})^2 - 2b(X - \bar{x})(Y - \bar{y}) + b^2(X - \bar{x})^2 + 2\big((Y - \bar{y}) - b(X - \bar{x})\big)(\bar{y} - a - b\bar{x})\big)^2 + (\bar{y} - a - b\bar{x})^2\Big) \\
&= V(Y) - 2b\,\mathrm{Cov}(X,Y) + b^2V(X) + (\bar{y} - a - b\bar{x})^2 \\
&= V(X)\left(b - \frac{\mathrm{Cov}(X,Y)}{V(X)}\right)^2 + \Big(E(Y) - a - bE(X)\Big)^2 + V(Y) - \frac{\mathrm{Cov}(X,Y)^2}{V(X)}
\end{aligned}$$

ここで，2 行目から 3 行目への変形では $E\big((X - \bar{x})^2\big) = V(X)$，$E\big((Y - \bar{y})^2\big) = V(Y)$，$E\big((X - \bar{x})(Y - \bar{y})\big) = \mathrm{Cov}(X,Y)$，$E(X - \bar{x}) = E(Y - \bar{y}) = 0$ を用いた．最後の行は a，b の 2 次式であり，その値が最小になるのは $(\quad)^2$ の中がともに 0 になるときである．それより直ちに (3) が得られる．