4. 回帰直線

1. 回帰分析とは

2 変量については，次の 2 つが主な考察の対象になる。

(1) 2 変量の間の相関関係（相関係数）
(2) 一方の変量から他方の変量の値を予測（回帰直線）

(1)で相関が認められた場合，一方の変量が他方の変量を左右または決定していると思われる場合がある。例えば，日本人全体を 1 つの集団とし，

$$y = \text{血圧},\quad x = \text{年齢}$$

と置けば，x と y の間には正の相関があることが知られている。この場合，年齢が高くなるから血圧も一般的に高くなるのであって，血圧が高くなるから年齢が高くなるわけではない。年齢が血圧をある程度決定している，と考えるのが妥当であろう。

このように，変量の間には，原因と結果という因果関係が考えられる場合がある。上の例では，年齢が原因で，血圧が結果である。

一般に，変量を「原因となる変量 x 」と「結果となる変量 y 」に分類して，それぞれを説明変数，被説明変数呼ぶ。(他にも次のような呼び方がある。)

○ 結果となる変量 y … 被説明変数（従属変数，目的変数）
○ 原因となる変量 x … 説明変数（独立変数）

回帰分析とは，このような x と y について，それらの間で成立する方程式を考察することだが，通常は，最小 2 乗法による回帰直線 $y = ax + b$ を考える。回帰直線を求めれば，x の値から y の値を予測・推定できることになる。

なお，x が被説明変数，y が説明変数の場合は，回帰直線 $x = ay + b$ を考えることになるが，話を簡単にするために，$y = ax + b$ の場合のみを説明する。

（注）血圧を左右する要因は，年齢の他に「性別」「食塩の摂取量」など，いろいろあるだろう。

$$y = \text{血圧},\quad x_1 = \text{年齢},\quad x_2 = \text{性別},\quad x_3 = \text{食塩の摂取量}$$

と置き，y を被説明変数，x_1, x_2, x_3 を説明変数と考えることもできる。

一般に，「1 つの被説明変数」と「複数の説明変数」との関係を考察する分野が「重回帰分析」であり，これは多変量解析の 1 つの分野である。説明変数が 1 つの場合を「単回帰分析」と呼ぶ。

重回帰分析では，変数の関係を

$$y = a_1 x_1 + a_2 x_2 + a_3 x_3 + b$$

のような 1 次の方程式で考察する場合が多い。この方程式による分析を，「線形回帰分析」と呼ぶ。

以下に述べる話は，説明変数が 1 つの線形回帰である。

2. 回帰直線の考え方

2 つの変量 x と y を観測（測定）して，2 次元データ (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) を得たとする。

x と y の間に強い相関があるとき，各点 (x_i, y_i) は 1 つの直線 $y = ax + b$ のまわりに集まっていると考えられる。簡単に言えば，この直線が回帰直線である。ただし，回帰直線は，相関とは無関係に定義できる。回帰直線の考え方は，以下の通りである。

(1) x の値から y の値を予測するための最適な直線を考える。最適とは何か？ 考え方はいろいろあるが，ここでは，最小 2 乗法による解を最適な直線とする。

(2) それを求めるために，まず直線 $y = ax + b$ を考える。（この段階では，a，b は定数である。）

(3) 当然，すべての点 (x_i, y_i) がこの直線上にあるわけではないので，一般には

$$y_i \neq ax_i + b$$

である。つまり，次の意味での実測値と予測値は一致しない。

① y_i … 実測値（観測した値）
② $ax_i + b$ … 予測値（直線から計算した値）

(4) ①と②の違いを「誤差」と解釈し，次のように置く。

誤差 : $e_i = y_i - (ax_i + b)$
誤差の 2 乗の和 : $f(a, b) = \sum_i e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$

一般に，$e_i \neq 0$ で，$f(a, b) \geq 0$ である。

(5) 一般に $e_i \neq 0$ だが，a, b を動かせば，各誤差を平均的に小さくすることはできる。つまり，$f(a, b)$ を a と b の関数ととらえ，その関数の値が最小になるような a と b の値は決定できる。

(6) 一般に，「2 乗の和を最小にするような解」を最適合の解とする方法を，「最小 2 乗法」と呼ぶ。

(7) ここでは，誤差の 2 乗の和 $f(a, b)$ が最小になるような a, b の値を求め，そのときの直線 $y = ax + b$ を「最適な直線」と考える。これが回帰直線である。

● 回帰直線の定義

変量 x と y から得られた 2 次元データ (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) に対して，a，b の関数

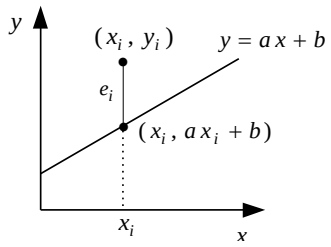
$$f(a, b) = \sum_{i=1}^{n} \{ y_i - (ax_i + b) \}^2$$

が最小になるように a，b の値を定める。

このとき，直線 $y = ax + b$ を「y の x への回帰直線」，傾き a を「y の x への回帰係数」という。

図形的には，次の意味になる。

y の x への回帰直線とは，各点 (x_i, y_i) から直線 $y = ax + b$ までの y 軸方向の距離の 2 乗の和が最小になる直線のことである



y の x への回帰直線

$e_i = y_i - (ax_i + b)$

距離の2乗の和が $\sum_i e_i^2$ が最小

上記の(4)においては，次のことに注意しておこう。

すべての点 (x_i, y_i) が，ある直線 $y = ax + b$ 上にあれば， $f(a, b) = 0$ となるので，この直線が回帰直線になる。逆に，ある実数 a, b に対して $f(a, b) = 0$ ならば，直線 $y = ax + b$ が回帰直線になり，すべての点 (x_i, y_i) はこの直線上にある。

回帰直線の式を求めると，次のようになる。

● 回帰直線の方程式

変量 x と y から得られた 2 次元データ (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) について，x_i の平均を $\bar{x}$，y_i の平均を $\bar{y}$ とする。

(1) y の x への回帰直線の式は

$$y - \bar{y} = \frac{\sigma_y}{\sigma_x} r_{xy} (x - \bar{x})$$

(2) y の x への回帰係数 a は，次の式で計算できる。

$$a = \frac{\sigma_y}{\sigma_x} r_{xy} = \frac{\sigma_{xy}}{{\sigma_x}^2} = \frac{n \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n \sum_i {x_i}^2 - (\sum_i x_i)^2}$$

上記を「偏微分」を使って証明している本も多いが，以下のように，高校の数 I の範囲でも証明できる。

a, b の関数 $f(a, b)$ が最小値をとるときの a, b の値を求めればよいが，常に不等式が

$$f(a, b) \geq k \quad (k \text{ は } a, b \text{ に無関係な定数})$$

が成立し，さらに等号が成立する場合があれば， k は $f(a, b)$ の最小値になることに注意しよう。

（証明） $e_i = y_i - (ax_i + b)$ とおくとき， a, b の関数

$$f(a, b) = \sum_i {e_i}^2$$

が最小になるようなa, bの値を求めればよい。そこで,

$$e_i = y_i - a x_i - b = (y_i - \bar{y}) - a(x_i - \bar{x}) + (\bar{y} - a\bar{x} - b)$$

と変形し, $\bar{y} - a\bar{x} - b = c$とおくと

$$e_i = y_i - a x_i - b = (y_i - \bar{y}) - a(x_i - \bar{x}) + c$$

これを2乗すると,

$$e_i^2 = (y_i - \bar{y})^2 + a^2(x_i - \bar{x})^2 + c^2 - 2a(x_i - \bar{x})(y_i - \bar{y}) - 2ac(x_i - \bar{x}) + 2c(y_i - \bar{y})$$

ここで, 偏差の総和は0だから, $\sum_i (x_i - \bar{x}) = 0$, $\sum_i (y_i - \bar{y}) = 0$

$$\therefore \quad \sum_i e_i^2 = \sum_i (y_i - \bar{y})^2 + a^2 \sum_i (x_i - \bar{x})^2 + nc^2 - 2a \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

さらに,

分散の定義から, $\frac{1}{n}\sum_i (x_i - \bar{x})^2 = \sigma_x^2$, $\frac{1}{n}\sum_i (y_i - \bar{y})^2 = \sigma_y^2$

相関係数の定義から, $\sigma_{xy} = \sigma_x \sigma_y r_{xy}$

共分散の定義から, $\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sigma_{xy} = \sigma_x \sigma_y r_{xy}$

よって,

$$\begin{aligned}
\frac{1}{n} f(a,b) &= \frac{1}{n}\sum_i e_i^2 \\
&= \frac{1}{n}\sum_i (y_i - \bar{y})^2 + a^2 \frac{1}{n}\sum_i (x_i - \bar{x})^2 + c^2 - 2a\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \\
&= \sigma_y^2 + a^2\sigma_x^2 + c^2 - 2a\sigma_x\sigma_y r_{xy} \\
&\geq \sigma_y^2 + a^2\sigma_x^2 - 2a\sigma_x\sigma_y r_{xy} \quad (c = \bar{y} - a\bar{x} - b = 0 \text{ のとき等号成立}) \\
&= (a^2\sigma_x^2 - 2a\sigma_x\sigma_y r_{xy}) + \sigma_y^2 \\
&= (a\sigma_x - \sigma_y r_{xy})^2 - \sigma_y^2 r_{xy}^2 + \sigma_y^2 \\
&= (a\sigma_x - \sigma_y r_{xy})^2 + \sigma_y^2 (1 - r_{xy}^2) \\
&\geq \sigma_y^2 (1 - r_{xy}^2) \quad (a\sigma_x - \sigma_y r_{xy} = 0 \text{ のとき等号成立})
\end{aligned}$$

$$\therefore \quad \frac{1}{n} f(a,b) \geq \sigma_y^2 (1 - r_{xy}^2)$$

従って, 常に

$$f(a,b) \geq n\sigma_y^2 (1 - r_{xy}^2)$$

が成立し, 等号が成立するのは

$$a\sigma_x - \sigma_y r_{xy} = 0 \quad \text{かつ} \quad \bar{y} - a\bar{x} - b = 0$$

の場合，すなわち

$$a = \frac{\sigma_y}{\sigma_x} r_{xy} \quad \text{かつ} \quad b = \bar{y} - a\bar{x}$$

のときである。従って，このとき $f(a, b)$ は最小になり，最小値は $n\sigma_y{}^2 (1 - r_{xy}{}^2)$ である。

よって，y の x への回帰直線は

$$y = ax + b = ax + \bar{y} - a\bar{x} \quad \text{すなわち} \quad y - \bar{y} = a(x - \bar{x})$$

であり，回帰係数 a は上式で与えられる。

また，$a = \frac{\sigma_y}{\sigma_x} r_{xy} = \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x{}^2}$ であり，

共分散の公式から，$\sigma_{xy} = \frac{1}{n} \sum_i x_i y_i - \bar{x}\bar{y}$

分散の公式から，$\sigma_x{}^2 = \frac{1}{n} \sum_i x_i{}^2 - \bar{x}^2$

平均の定義から，$n\bar{x} = \sum_i x_i$ ，$n\bar{y} = \sum_i y_i$

であるから，a を求める式は次のように変形できる。以上で証明された。

$$a = \frac{n^2 \sigma_{xy}}{n^2 \sigma_x{}^2} = \frac{n \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n \sum_i x_i{}^2 - (\sum_i x_i)^2}$$

3. 相関係数と回帰直線との関係

y の x への回帰直線を $y = ax + b$ とすると，以下が成り立つ。

(1) 回帰直線 $y = ax + b$ は，次の式で与えられた。

$$y - \bar{y} = a(x - \bar{x}), \quad a = \frac{\sigma_y}{\sigma_x} r_{xy}$$

(2) すべての点 (x_i , y_i) が直線上にあれば，その直線が回帰直線であり，$r_{xy} = \pm 1$ である。逆に，$r_{xy} = \pm 1$ であれば，すべての点 (x_i , y_i) は回帰直線上にある。

(3) 次の関係があることがわかる。

$r_{xy} > 0 \quad \Leftrightarrow$ 回帰直線の傾き a は正
$r_{xy} < 0 \quad \Leftrightarrow$ 回帰直線の傾き a は負

(4) 回帰直線は，平均を座標とする点 $(\bar{x}, \bar{y})$ を通る。すなわち，

$$\bar{y} = a\bar{x} + b$$

が成り立つ。

(5) 各点(x_i, y_i)について，次を考える。

- y_i … 実測値（観測した値）
- $ax_i + b$ … 予測値（回帰直線から決まる値）
- 誤差（残差） $e_i = y_i - (ax_i + b)$

さらに，誤差e_iの分散

$$\sigma_e{}^2 = \frac{1}{n}\sum (e_i - \bar{e})^2$$

を考えると，以下が成り立つ。（誤差e_iの分散を「誤差分散」「残差分散」という。）

① 予測値の平均 ＝ 実測値の平均

なぜなら，予測値$ax_i + b$の平均は$a\bar{x} + b$，実測値y_iの平均は$\bar{y}$である。よって，(4)より自明である。

② 予測値の分散は，$a^2\sigma_x{}^2 = r_{xy}{}^2 \cdot \sigma_y{}^2$

なぜなら，予測値$ax_i + b$の分散は$a^2\sigma_x{}^2$である。一方，(1)より，$a\sigma_x = r_{xy} \cdot \sigma_y$である。

③ 誤差の合計は0，誤差の平均は0である。すなわち

$$\sum_i e_i = 0, \quad \bar{e} = 0$$

なぜなら，①より，予測値の合計と実測値の合計は等しい。従って，

$$\sum_i e_i = \sum_i \{y_i - (ax_i + b)\} = \sum_i y_i - \sum_i (ax_i + b) = 0$$

④ 誤差分散$\sigma_e{}^2$は，次の式で与えられる。

$$\sigma_e{}^2 = \frac{1}{n}\sum e_i{}^2 = \sigma_y{}^2(1 - r_{xy}{}^2)$$

なぜなら，③より，前半の等式は明らか。また，後半の等式は，次の等式から自明。

$$f(a, b) = \sum_i e_i{}^2 = n\sigma_y{}^2(1 - r_{xy}{}^2)$$

(4) 誤差分散は

$$\sigma_e{}^2 = \frac{1}{n}\sum e_i{}^2 = \sigma_y{}^2(1 - r_{xy}{}^2)$$

になった。この等式からでも，$1 - r_{xy}{}^2 \geq 0$，すなわち，$-1 \leq r_{xy} \leq 1$であることがわかる。

また，この等式から，r_{xy}の値が1や−1に近くなるほど$1 - r_{xy}{}^2$は小さくなるので，分散$\sigma_e{}^2$の値は小さくなり，各点は回帰直線に近づいていくことがわかる。同様に，r_{xy}が0に近づくほど$1 - r_{xy}{}^2$は大きくなるので，$\sigma_e{}^2$の値は大きくなり，各点は回帰直線から離れていく。

すなわち，r_{xy}の値が± 1に近いほど，誤差e_iは全体的に小さくなっていくので，回帰直線による予測値は信頼できる値になっていく。

● **注意**

ある変量の分散の値が小さくなるということは，変量のとる値のバラツキが小さくなるということ，すなわち，変量のとる値が平均に集中していくことである。言い換えれば，変量のとる値と平均とのズレ（偏差）が小さくなっていくことである。逆に，分散が大きくなるということは，変量のとる値と平均とのズレが大きくなっていくことである。

誤差e_iの平均は 0 であるから，分散$\sigma_e{}^2$が小さくなればなるほど，e_iの値は平均 0 に近づくので，予測値は実測値に近づいていく。この実測値には，観測した値のみならず，将来観測される値も含まれるので，回帰直線による予測が可能になる。

4. 回帰直線の計算

回帰直線の求め方はいろいろあるが，前述の式で計算してみよう。すべての点(x_i, y_i)が表のように与えられているとき，y の x への回帰直線$y = ax + b$を求めるわけである。

表において，x_i と y_i の値から，$x_i{}^2$ と $x_i\ y_i$ の値を求めておく。

x_i	y_i	$x_i{}^2$	$x_i\ y_i$
7	9	49	63
5	5	25	25
8	8	64	64
4	5	16	20
7	6	49	42
6	7	36	42
2	4	4	8
6	8	36	48
5	8	25	40
10	10	100	100
60	70	404	452

（最下行は合計）

$n = 10$ であり，平均は

$$\bar{x} = 60/10 = 6,\quad \bar{y} = 70/10 = 7$$

回帰係数 a を求めると，

$$a = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n\sum_i x_i{}^2 - (\sum_i x_i)^2}$$

$$= \frac{10 \times 452 - 60 \times 70}{10 \times 404 - 60^2}$$

$$= \frac{4520 - 4200}{4040 - 3600} = \frac{320}{440} = 0.73$$

よって，y の x への回帰直線の式は，

$$y - 7 = 0.73\,(x - 6)$$

5. 練習問題

回帰分析は様々な分野で利用されるが，企業経営では不可欠な分析手法である。簡単な例で考えてみよう。

次のデータは，A 社の過去 5 年間の「セールスマン数」と「売上高」のデータである。

	セールスマン数（人）	売上（千万円）
2011 年度	1	3
2012 年度	2	3
2013 年度	3	4
2014 年度	4	5
2015 年度	5	5
2016 年度		

● 問題１

セールスマン数と売上との間には，相関はあるのか。

● 問題２

もし強い正の相関があれば，2016 年度は思い切ってセールスマン数を 2 人増やし「7 人」にすることにする。その場合，どれほどの売上が見込まれるだろうか。

＜考え方＞

(1) セールスマン数 ＝x，売上 ＝y とおき，散布図を描く。いきなり相関係数を計算せずに，まずグラフで判断する。相関以外の関係や外れ値が読み取れるかもしれないからである。

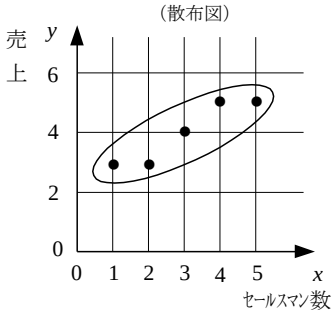
(2) 相関があると判断できた場合，相関係数 r_{xy} を計算し，何らかの基準で「強い」「弱い」を判断する。ここでは，次の基準で判断してみる。

判断基準： $|r_{xy}| > 0.7$ … 強い相関

(3) 強い相関の場合，y の x への回帰直線 $y = ax + b$ を求め，$x = 7$ のときの y の値（予測値）を計算する。

＜解答＞

(1) 散布図から，正の相関があると判断できる。



(2) 次に，x と y の相関係数を r_{xy} を求める。

分散の公式 $\sigma_x{}^2 = \frac{1}{n}\sum_i x_i{}^2 - \bar{x}^2$

共分散の公式 $\sigma_{xy} = \frac{1}{n}\sum_i x_i y_i - \bar{x}\,\bar{y}$

相関係数 $r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

x_i	y_i	$x_i{}^2$	$y_i{}^2$	$x_i y_i$
1	3	1	9	3
2	3	4	9	6
3	4	9	16	12
4	5	16	25	20
5	5	25	25	25
15	20	55	84	66

（最下行は合計）

変量 x ：平均 $\bar{x} = \dfrac{15}{5} = 3$，分散 $\sigma_x{}^2 = \dfrac{55}{5} - 3^2 = 11 - 9 = 2$

変量 y ：平均 $\bar{y} = \dfrac{20}{5} = 4$，分散 $\sigma_y{}^2 = \dfrac{84}{5} - 4^2 = 16.8 - 16 = 0.8$

共分散 $\sigma_{xy} = \dfrac{66}{5} - 3 \times 4 = 13.2 - 12 = 1.2$

相関係数 $r_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = \dfrac{1.2}{\sqrt{2} \times \sqrt{0.8}} \fallingdotseq 0.949$

よって，$r_{xy} > 0.7$ であるから，強い正の相関があると判断できる。

(3) y の x への回帰係数 a は

$$a = \frac{\sigma_y}{\sigma_x} r_{xy} = \frac{\sigma_y}{\sigma_x} \times \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x{}^2} = \frac{1.2}{2} = 0.6$$

よって，y の x への回帰直線は

$$y - 4 = 0.6(x - 3) \quad \therefore \quad y = 0.6x + 2.2$$

ここで，$x = 7$ のとき，

$$y = 0.6 \times 7 + 2.2 = 6.4$$

よって，2016 年度にセールスマン数を 7 人にすれば，6 千 4 百万円の売上が見込まれる。

（注意）(2)のように相関係数を求めれば，回帰係数はすぐに計算できる。求めた相関係数の値が 1 に非常に近いので，上の予想額は信頼できる値である。ただし，実際には，相関係数・回帰係数の推定や検定を行い，より厳密な分析をする。

6. 「回帰」の名前の由来

有名な話なので紹介しておこう。

親の身長と子どもの身長の間には，正の相関があることは確かである。身長の高い親の子どもは一般に背が高く，身長の低い親の子どもは一般に背が低い。しかし，子どもの身長は，親の身長と完全に同じということはない。親よりも背が高い子どももいれば，低い子どももいる。言い換えれば，子どもの身長は，親の身長の上下にばらついているわけである。

では，子どもの身長は，常に親の身長の上下にばらつくのだろうか。もしそうであれば，大変なことになる。身長の高い親からは更に身長が高い子どもが生まれ，身長の低い親からは更に身長が低い子どもが生まれことになり，人類が誕生してから数百万年の世代交代を考えれば，現在の「人間の身長」が説明できなくなる。

英国の遺伝学者ゴールトン（Francis Galton，1822-1911）は，父親の身長（x）が，成人した息子の身長（y）にどのように遺伝するかを研究し，ロンドンで多くの親子の身長を測定した。得られたデータの散布図を描くと，正の相関が見られた。さらに，父親の身長 x を 1 インチごとの階

級に分割し，この階級に入る y の平均値を求めて点をプロットすると，ほぼ次の一直線になった。

$$y = 0.516x + 33.73$$

この式は，x の値が 1 インチ増えると，y の値は 0.516 インチ増えることを意味する。

ゴールトンは，この直線の傾きが 0.516 で 1 より小さいこと，y 切片が 33.73 であることに注目して，「背の高い父親の息子は，その父親ほどには高くなく，背の低い父親の息子は，その父親ほどには低くない」ということを発見した。そして，この現象を「平均への回帰（regression）」と名づけた。

つまり，父親の身長が父親の集団平均値から離れていくと，息子の身長も息子の集団平均値から離れていくが，父親の身長がその平均値から大きく離れていくと，息子の身長は逆にその平均値に戻っていくという現象である。父親の集団平均値と息子の集団平均値はほぼ等しいので，息子の身長が父親の集団平均値に帰っていくわけである。ゴールトンは，これを“先祖がえり”の現象と考え，回帰と名付けたのである。

先祖がえりしなければ，成人の身長の最大値は際限なく増大し，最小値は際限なく減少するわけであるから，この回帰現象も当然ともいえる。上記は回帰直線であり，その傾きが 1 ではなく，約 0.5 ということから，回帰現象が起こる。

なお，相関という言葉をはじめて用い，相関係数を r で表したのもゴールトンである。