第 2 章　相関と単回帰

1　散布図と相関

散布図の例：**図 2-1**（9 ページ）の日本のマクロ経済時系列データに関する散布図

横軸：1980 年から 1995 年までの日本の家計消費支出（名目，年次データ）

縦軸：家計消費支出の主要な 4 つの構成項目

Ａ．食品・飲料・煙草　　Ｂ．家賃・水道・光熱

Ｃ．医療・保健　　Ｄ．レクリエーション・娯楽・教育・文化サービス

に関する名目値の年次データ

データ：**表 2-1**（「国民経済計算年報」（経済企画庁）の四半期データを変換）

散布図：2 つの変数間の関係を視覚的に捉えることができる．

→　家計消費と構成項目のそれぞれとは右上がりの関係があることが鮮明に読みとれる．

相関係数：2 つの変数の直線的な関係の強さを具体的な数値で表したもの

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{s_{xy}}{s_x s_y}\qquad\text{（無名数）}\qquad(1)$$

s_x と s_y はそれぞれ x と y の**標準偏差**であり，

$$s_x=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\ ,\qquad s_y=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

で定義される．さらに，

$$s_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\quad：\ x\ と\ y\ の\textbf{共分散}$$

相関係数 r_{xy} の性質

(a) 常に $|r_{xy}|\le 1$ である．

(b) 特別の値 1（正の完全相関）と -1（負の完全相関）になるのは次の場合である．

$r_{xy}=1$　　すべてのデータが右上がりの直線 $y=\bar{y}+\frac{s_y}{s_x}(x-\bar{x})$ 上にある．

$r_{xy}=-1$　　すべてのデータが右下がりの直線 $y=\bar{y}-\frac{s_y}{s_x}(x-\bar{x})$ 上にある．

(c) x のデータを a 倍して b を加えたデータと，y のデータを c 倍して d を加えたデータの相関係数は，$ac>0$ ならば r_{xy} であり，$ac<0$ ならば $-r_{xy}$ である．

〔**例題 2.1**〕**表 2-1** から，家計消費と他の 4 種類のデータとの相関係数を計算せよ．

（解）家計消費とA（食品・飲料・煙草），B（家賃・水道・光熱），C（医療・保健），D（レクリエーション・娯楽・教育・文化サービス）の間には，**図 2-1** の散布図から強い正の相関が示唆される．特に，Dとの相関がもっとも強いものと予想される．実際，それぞれの相関係数は次のようになる．

	A	B	C	D
家計消費	0.985	0.986	0.984	0.998

D以外との 3 つの相関係数も 1 にかなり近い値で，同程度の値となっている．一般に，散布図から受ける印象だけで相関係数の値を推論することは困難である．逆に，相関係数が同じであるからといって，散布図の形状が同じになるわけではないことに注意されたい．

経済時系列のデータでは，時間的な要因が介在して**見せかけの相関**が生み出される場合がある．

〔**例題 2.2**〕変数 x と y の時系列データが，時間を横軸とする別々の右上がりの直線上にあるとする．このとき，相関係数は 1 となることを示せ．

（解）時点 t における x と y のデータは

$$x_t = a + bt\,, \qquad y_t = c + dt$$

と表すことができる．ここで，b と d は正である．このことから，

$$t = \frac{x_t - a}{b} = \frac{y_t - c}{d}$$

となるので，

$$y_t = c - \frac{ad}{b} + \frac{d}{b} x_t$$

を得る．ここで，d/b は正であるから，x と y のデータは右上がりの直線上にあることになる．したがって，相関係数は 1 となる．

2　回帰の意味

変数 x と y の間に

$$y = \alpha + \beta x \tag{2}$$

という線形関係を想定する．このことは，平面上に散らばっている点を直線上に縮約することを意味する．

このように，ある変数を他の変数の関数（今の場合は線形関数）として表現して縮約することを**回帰**という．特に，(2) のように単一の変数の関数として表現される場合を**単回帰**という．右辺の変数 x は**説明変数**（あるいは**独立変数**），左辺の変数 y は**被説明変数**（あるいは**従属変数**）と呼ばれる．また，α は切片，β は傾きである．β は，x の 1 単位の増加に対する y の増加を表す**限界性向**である．

回帰式 (2) を想定したときの統計的問題は，切片 α，傾き β をどのように求める

かということである．

回帰の本来の意味　→　図 2-2 参照

3　最小 2 乗法

データ（x_i，y_i）（$i=1,\cdots,n$）が与えられたとき，回帰直線の切片と傾きを求めるための方法はいくつかある．最もよく使われる方法は

$$S(\alpha,\beta)=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 \tag{3}$$

を最小にするように α と β を決めることである．すなわち，**図 2-3** において，データ y_i（$i=1,\cdots,n$）と直線上の値との縦軸方向の差を最小にするような直線を求めるわけである．このようにして未知の値を求める方法を**最小 2 乗法**という．

最小 2 乗法から得られる α と β の値をそれぞれ a，b とすれば，それらは (3) 式の $S(\alpha,\beta)$ を α と β に関して偏微分して 0 とおいた方程式の解である．すなわち，a と b は

$$a\sum_{i=1}^{n}1+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \tag{4}$$

$$a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i \tag{5}$$

という連立方程式の解となることがわかる．この連立方程式は**正規方程式**と呼ばれる．

正規方程式を解くことにより，

$$a=\bar{y}-b\bar{x} \tag{6}$$

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{s_{xy}}{s_x^2}=r_{xy}\frac{s_y}{s_x} \tag{7}$$

が得られる．

4　回帰直線の経済学的解釈

回帰直線が求まれば，経済学に関連する次のような特性値を計算することができる．ただし，ここでは右上がりの回帰直線を考え，x も y もともに正の値をとるものとする．

回帰直線 $y=a+bx$ から得られる経済学的特性値

(a)	**限界性向**	$\dfrac{dy}{dx}=b$	x が 1 単位増加したときの y の増加量
(b)	**平均性向**	$\dfrac{y}{x}=\dfrac{a}{x}+b$	x に占める y の割合
(c)	**弾力性**	$\dfrac{dy/y}{dx/x}=\dfrac{bx}{y}=\dfrac{bx}{a+bx}$	x が 1% 増加するときの y の % 増加率

平均性向と弾力性は，変数 x あるいは y に依存するので，実際の計算においてはそれぞれの平均でおきかえるのが普通である．また，経済データで普通に見られるような状況，すなわち，回帰直線が右上がりで，x も y もともに正の値をとる場合には，平均性向は，切片が正ならば減少し，負ならば増加する．また，弾力性は，同じ状況で考えると，切片が負ならば弾力的，正ならば非弾力的になることがわかる．

〔**例題 2.3**〕**表 2-1** のデータにおいて，4 つの消費項目のそれぞれを家計消費で説明する回帰直線を求め，限界性向，平均性向，弾力性を計算せよ．

（解）例題 2.1 のように，4 つの消費項目をA（食品・飲料・煙草），B（家賃・水道・光熱），C（医療・保健），D（レクリエーション・娯楽・教育・文化サービス）で表し，それぞれの消費額を a，b，c，d とする．また，家計消費を x とすると，$\bar{x} = 214.7125$ であり，回帰直線などは以下のように求められる．ここで，平均は左辺の変数のデータの平均である．また，平均性向と弾力性は，平均で評価されている．

	回帰直線	限界性向	平均	平均性向	弾力性
A：	$a = 21.54 + 0.101x$	0･101	43.15	0.201	0.498
B：	$b = -10.84 + 0.246x$	0.246	42.08	0.196	1.258
C：	$c = 0.828 + 0.0973x$	0.097	21.72	0.101	0.962
D：	$d = -13.30 + 0.176x$	0.176	24.39	0.114	1.546

消費項目Aは，平均性向は高いものの，限界性向は低く，非弾力的である．BとDは，平均性向は低いものの，限界性向が高く，切片が負であることから，弾力的である．Cは限界性向も平均性向もともに低く，弾力性の程度も低くなっている．

上で扱ったのは時系列データであるが，同様の分析は横断面データに対しても行うことができる．例として，表 2-2 の家計の消費支出に関するデータを考えよう．これは，総務庁統計局から公表されている「家計調査年報」から抜粋したものであるが，データは収入の大きさで同じ度数の 10 個の階級ごとにまとめられており，数値は各階級の平均である．

表 2-2 年間収入十分位階級別 1 世帯当りの月間消費支出と構成項目
（勤労者世帯，1996年，単位：万円）

〔**例題 2.3**〕**表 2-2** の十分位階級別データから，家計の月間消費と 3 つの構成項目（食費，教養・娯楽費，直接税）の相関係数を求めよ．また，構成項目の各々を月間消費で説明する回帰直線を求め，限界性向，平均性向，弾力性を計算せよ．

（解）まず，月間消費と3 つの構成項目との相関係数は次の通りである．

	食費	教養・娯楽費	直接税
月間消費	0.983	0.986	0.968

次に，3 つの消費項目をA（食費），B（教養・娯楽費），C（直接税）で表し，それぞれの支出額を a,b,c（単位：万円） とする．また，月間消費を x とすると，$\bar{x}=35.175$ であり，回帰直線などは例題 2.2 と同様にして，以下のように求められる．

	回帰直線	限界性向	平均	平均性向	弾力性
A：	$a=2.783+0.143x$	0･143	7.813	0.222	0.644
B：	$b=-0.769+0.118x$	0.118	3.381	0.096	1.228
C：	$c=-7.658+0.347x$	0.347	4.561	0.130	2.676

BとCについては，切片が負であることから，弾力性が 1 より大きくなる．特に，C（直接税）に関しては，弾力性が 2.676 と非常に大きな値となっている．なお，直接税については，収入に対する回帰直線を求めることにより，累進課税の様子がよくわかるであろう．

5　回帰直線の性質

回帰直線の切片と傾きは，正規方程式 (4)，(5) を解くことにより，(6)，(7) のように求められるが，さらに，回帰直線に関する性質を調べるために，

$$\hat{y}_i=a+bx_i \quad ；\text{実績値 } y_i \text{ の}\textbf{内挿値}（\textbf{理論値}） \tag{8}$$

$$e_i=y_i-\hat{y}_i \quad ；\text{ 回帰の}\textbf{残差}（\textbf{実績値－理論値}） \tag{9}$$

を定義しよう．このとき，正規方程式 (4)，(5) は次のように書き換えられる．

$$\sum_{i=1}^{n}\{y_i-(a+bx_i)\}=\sum_{i=1}^{n}(y_i-\hat{y}_i)=\sum_{i=1}^{n}e_i=0 \tag{10}$$

$$\sum_{i=1}^{n}x_i\{y_i-(a+bx_i)\}=\sum_{i=1}^{n}x_i(y_i-\hat{y}_i)=\sum_{i=1}^{n}x_ie_i=0 \tag{11}$$

(10) から，残差の和は 0，したがって，残差の平均も 0 である．また，(11) から，説明変数と残差は直交する．これらのことから，次の分解が成り立つ．

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2 \tag{12}$$

全変動 ＝ 回帰変動 ＋ 残差変動

全変動の分解により，回帰直線のあてはまりの程度を測る指標として，

$$\textbf{決定係数} \;=\; R^2 \;=\; \text{回帰変動／全変動} \tag{13}$$

を考えることができる．決定係数は明らかに 0 と 1 の間の数であり，1 に近いほどあてはまりがよいと考えられる．

なお，最後に述べた決定係数については，単回帰の場合には相関係数と次のような関係にある．

〔**例題 2.4**〕 y を x に回帰して得られる単回帰の決定係数は，x と y の相関係数の 2 乗に等しいことを示せ．

（解）回帰変動は次のように書き換えることができる．

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}\{a+bx_i-(a+b\bar{x})\}^2=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= r_{xy}^2 \frac{s_y^2}{s_x^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = r_{xy}^2 \times (n-1) s_y^2$$

最後の表現を全変動 $(n-1)s_y^2$ で割ったものが決定係数であるが，それは明らかに r_{xy}^2 となる．

今まで説明してきた回帰直線に関連した性質は，次のようにまとめることができる．

回帰直線に関連した性質

(a) 平均の点（$\bar{x}, \bar{y}$）を通り，傾きが $b = r_{xy} \times s_y / s_x$ の直線である．

(b) 残差の総和は 0 となる．従って，残差の平均も 0 となる．

(c) 残差と説明変数は直交する．

(d) 決定係数は相関係数の 2 乗に等しい．

6　原点を通る回帰直線

この節では，切片を 0 に制約して得られる回帰直線を考える．この場合を別扱いして取り上げる理由は，前節で述べた通常の回帰直線の性質の中で必ずしも成立しないものが出てくるからである．

切片が 0 の回帰直線を求めるには，傾きだけを考えればよいから，最小 2 乗法の考え方を使って，

$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

を最小にする β を求めることになる．解を b_1 とすると，b_1 は $S(\beta)$ を β に関して微分して 0 とおくことにより，

$$b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad \Leftrightarrow \qquad b_1 = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 \tag{14}$$

となる．このときの回帰直線は $y = b_1 x$ となり，平均の点（$\bar{x}, \bar{y}$）を通るとは限らない．

ところで，(14) の関係は，

$$\sum_{i=1}^{n} x_i (y_i - b x_i) = \sum_{i=1}^{n} x_i e_i = 0$$

のようにも表すことができる．この場合の残差は $e_i = y_i - b_1 x_i$ であり，残差がみたす性質は説明変数と直交することだけである．したがって，残差の和も平均も必ずしも 0 にはならない．このことから，(12) の形の変動の分解公式も成り立たない．この場合には，

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

の形の分解が成り立つことがわかる．ただし，$\hat{y} = b_1 x$ である．そして，決定係数についても，

$$R^2 = \sum_{i=1}^{n} \hat{y}_i^2 \Big/ \sum_{i=1}^{n} y_i^2$$

により定義することになる.

切片を 0 に制約した回帰直線の性質をまとめると次のようになる.

切片を 0 に制約した回帰直線に関連した性質

(a) 原点を通り，傾きが

$$b_1 = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2$$

の直線である．平均の点（$\bar{x}, \bar{y}$）を通るとは限らない.

(b) 残差と説明変数は直交する．しかし，残差の総和は一般に 0 でない.

(c) 決定係数は

$$R^2 = \sum_{i=1}^{n} \hat{y}_i^2 \Big/ \sum_{i=1}^{n} y_i^2$$

で定義される．相関係数の 2 乗に等しくなる保証はない.

切片が 0 に制約された回帰直線の性質は，次章で説明する重回帰の場合にも同様に成立する．しかし，繁雑さを避けるため，本書では以後断らない限り，切片に制約をおかない場合だけを考えることにする.

7　偏相関

偏相関係数: 相関係数は 2 つの変数間の関係の強さを測る指標であるが，一般には，これらの変数は他の変数の影響を受けて変動している．そのような他の変数の影響を除去したあとに得られる 2 つの変数の間の純粋な相関係数

z の影響を除去したときの x と y の偏相関係数

x と y を，それぞれ z に回帰したときの残差の相関係数

一般には，x と y に影響を与える変数は複数個ありうる．その場合には，これらの複数個の変数に対して，次章で説明する重回帰から残差を求めて，同様に考えればよい．次の例題は，z として直線トレンドを考えた場合の偏相関係数に関するものである.

〔**例題 2.5**〕**表 2-1** の家計消費と 4 つの構成項目のそれぞれについて，時間的な直線トレンドを除去したあとの偏相関係数を求めよ.

（解）前と同様に，4 つの消費項目をA（食品・飲料・煙草），B（家賃・水道・光熱），C（医療・保健），D（レクリエーション・娯楽・教育・文化サービス）で表し，それぞれの消費額を a，b，c，d とする．また，家計消費を x とする．これら 5 つの変数を，時間 t = 1,2,・・・,16 に回帰したときの回帰直線は次の通りである.

$$a = 34.22 + 1.05t \qquad b = 20.02 + 2.60t \qquad c = 12.94 + 1.03t$$

$$d = 8.81 + 1.83t \qquad x = 125.59 + 10.48t$$

いずれも，切片が正の右上がりの直線トレンドをもっている．そして，残差から計算される偏相関係数は次のようになる．参考までに，例題 2-1 で求めた通常の相関係数も併記しておいた．

	相関係数	偏相関係数
Ａと家計消費	0.985	0.621
Ｂと家計消費	0.986	0.239
Ｃと家計消費	0.984	-0.482
Ｄと家計消費	0.998	0.936

家計消費は，Ｄとは時間的要因を除去したあとでも，なお強い正の相関をもつが，ＡとＢ，特にＢとは相関が大幅に減少している．そして，Ｃとは負の相関に転じている．**図 2-5**（25 ページ）には，残差間の散布図が示されている．上で求めた偏相関係数は，これらの残差間の相関係数である．