2008 年度前期　応用統計学　第２回

回帰分析 **(1)** － 共分散と相関，線形単回帰

今回から３回で，回帰分析について説明します．回帰分析は，例えば「緯度が高くなると気温が下がる」「緯度が高く，標高が高くなると，気温が下がる」というように，ある変量の変化が他の変量の変化によって説明できるというモデルを考えて，変量間の関係を記述する手法です．今回は，回帰分析以外にも多変量解析の基盤となる，変量間の関連を表現する「相関」の考え方と，ある変量の変化を他のひとつの変量の変化で説明し，それらの変量の間に１次関数の関係があると仮定する「線形単回帰」を説明します．

相関関係と散布図

第１回で説明したように，「各県について，人口と店の数」「日本の各都市について，緯度と年平均気温」などのように，各個体について２種類以上の項目からなるデータを集めたとき，この「項目」を変量といいます．

例えば，「人口と店の数」では，人口が多い町では店の数も多い傾向があるでしょうし，「緯度と気温」では，緯度が高くなると気温が低くなる傾向があるでしょう．これらはあくまで「傾向」であって，店の数が人口だけで決まったり，気温が緯度だけで決まるわけではありません．しかし，そのような傾向があるのは確かです．

このような，「変量どうしの，互いの増減の傾向の関係」を相関関係といいます．「人口と店の数」のように，「人口が多いと店の数も多い」という関係を正の相関関係といい，「緯度と気温」のように「緯度が高いと気温は低い」という関係を負の相関関係といいます．

相関関係を視覚的に表現する方法として，散布図があります．散布図は，２つの変量の組からなる各データを，縦横の軸にそれぞれの変量をとって平面上に配置したものです．例えば，表 1 に示す「日本の各都市の，緯度と年平均気温」のデータ[1]を，散布図に描くと，図 1 のようになります．このように，負の相関関係は，散布図上では右下がりの直線上にデータが分布するように表現されます．また，正の相関関係では右上がりの直線上に並ぶことになります．各データが散布図上でほぼ一直線上に乗っており，「変量どうしの，互いの増減の傾向」がはっきりしているとき，「強い相関がある」といいます．これに対し，各データが直線からばらついていて増減の傾向がはっきりしないときは，「弱い相関がある」といいます．

共分散と相関係数

相関関係の強い／弱いを数値で表すのが相関係数です．データが $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ の n 組であるとき，x と y との相関係数 r_{xy} は

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

で表されます．上の式の中央の部分で，分母は，x, y それぞれの標準偏差の積です．分子は，x, y のそれぞれの偏差を同時に平均したもので，共分散といいます．

[1] 日本列島大地図館（小学館）［理科年表より転載］より

地名	緯度（度）	気温（℃）
札幌	43.05	8.0
青森	40.82	9.6
秋田	39.72	11.0
仙台	38.27	11.9
福島	37.75	12.5
宇都宮	36.55	12.9
水戸	36.38	13.2
東京	35.68	15.3
新潟	37.92	13.1
長野	36.67	11.4
静岡	34.97	16.0
名古屋	35.17	14.9
大阪	34.68	16.2
鳥取	35.48	14.4
広島	34.40	15.0
高知	33.55	16.3
福岡	33.92	16.0
鹿児島	31.57	17.3
那覇	26.20	22.0

表 1: 日本の年の緯度と気温

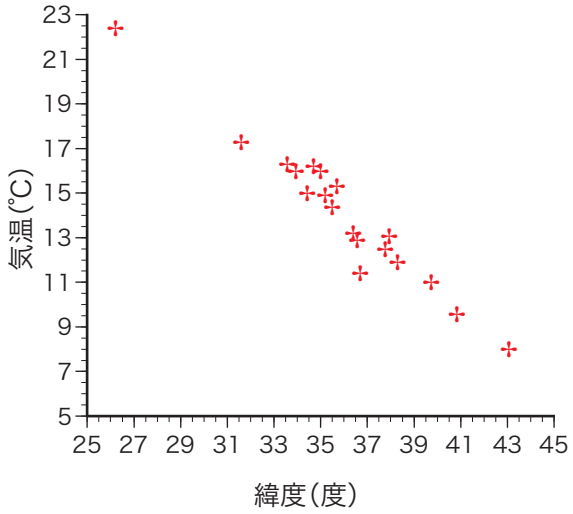

図 1: 散布図：緯度と気温の関係

共分散の意味を，図 2 で考えてみましょう．散布図の平面を，x の平均および y の平均を境にして四分割します．各領域で，$(x_i - \bar{x})(y_i - \bar{y})$ の値を考えてみます．

（イ）では，$x_i - \bar{x} > 0, y_i - \bar{y} > 0$ で，$(x_i - \bar{x})(y_i - \bar{y}) > 0$ であり，(x_i, y_i) が右上に行くほどこの積の値は大きくなります．また，（ハ）では $x_i - \bar{x} < 0, y_i ? \bar{y} < 0$ でやはり $(x_i - \bar{x})(y_i - \bar{y}) > 0$ であり，(x_i, y_i) が左下に行くほどこの積の値が大きくなります．これに対して，（ロ）や（ニ）では $(x_i - \bar{x})(y_i - \bar{y}) < 0$ となります．

では，図 3 の 3 つの分布で，$\sum_i (x_i - \bar{x})(y_i - \bar{y})$ の値はどうなるでしょうか？（グレーの部分にデータがおもに分布しているとします．）(a) の場合は先の図 2 の（イ）（ハ）の部分に多く分布していますから正の大きな値，(b) の場合は（ロ）（ニ）の部分に多く分布していますから負の大きな値，(c) の場合は（イ）（ロ）（ハ）（ニ）のすべての部分に分布しているので打ち消しあって 0 に近い値になります．

この $\sum_i (x_i - \bar{x})(y_i - \bar{y})$ を，グレーの部分に分布しているデータの個数 n に影響されないように，n で割って「合計」でなく「平均」にしたものが共分散です．つまり正の相関があるとき正の値，負の相関のとき負の値，どちらでもないときは 0 に近い値になります．

相関係数は共分散を x, y それぞれの標準偏差の積で割ったものとなっていますが，これは図 4 の左右の分布で相関係数が同じになるようにするためです．図 4 の左右は，ばらつきは異なっていますが，相関の強さは同じです．なお，相関係数は –1 から 1 の範囲の値をとり，1 がもっとも強い正の相関，–1 がもっとも強い負の相関，0 は相関がないことをあらわします．

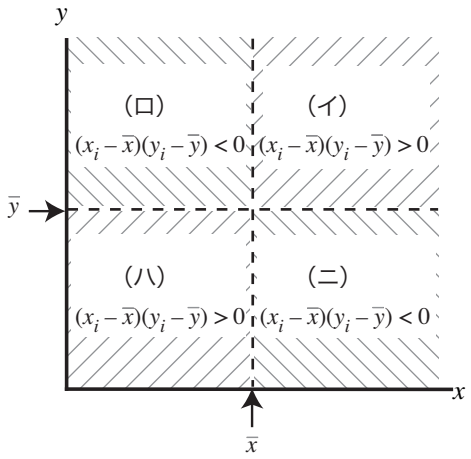

図 2: 共分散の概念

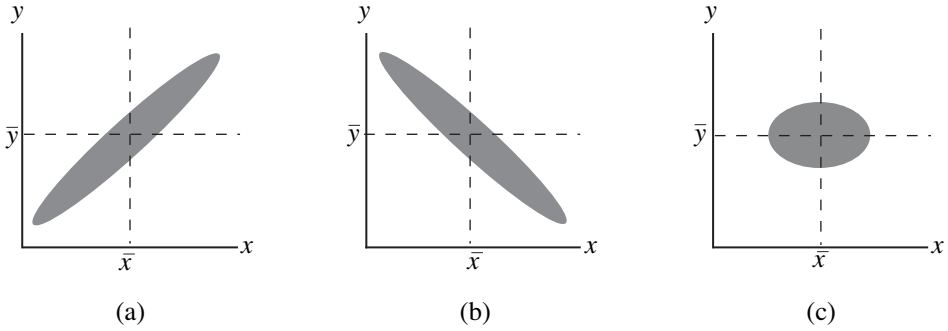

図 3: 正負の相関

線形単回帰

表 1・図 1 の緯度と気温の間には負の相関関係があります．そこで，相関関係にあるデータのばらつき方を，気温が緯度によって決まっているというモデルで表現しようというのが回帰分析です．緯度を x とし，気温を y とするとき，「x によって y が決まる」という関係になっていることを，統計学では「変量 y は変量 x によって説明される」といい，x を説明変量，y を被説明変量といいます．また，この関係を y の x 上への回帰といい，この例のように説明変量がひとつの場合を単回帰といいます．これに対し，説明変量が複数（たとえば緯度と標高）ある場合を重回帰といいます．

緯度 x と気温 y に散布図上で直線の関係があると仮定するということは，散布図上にばらついているデータを，$y = a + bx$ という式で表される直線というモデル，すなわち線形モデルで表すことになります．このような回帰を，線形回帰（この例の場合は，説明変量がひとつなので，線形単回帰）といいます．

そこで，この式の a, b つまりパラメータを決める方法を考えます．与えられている緯度と気温の組を (x_i, y_i) とします．x と y の間の関係が，$y = a + bx$ というモデルで完全に表されるのなら，$x = x_i$ のとき $y = a + bx_i$ となるはずです．しかし，現実には $y = y_i$ となっています．そこで，パラメータのさまざまな値のうちで，この「全ての (x_i, y_i) についての，y_i と $a + bx_i$ との差の合計」が，もっとも小さくなるパラメータをもっとも適切なパラメータとします．差には正負がありますから，実際には差の 2 乗の合計，

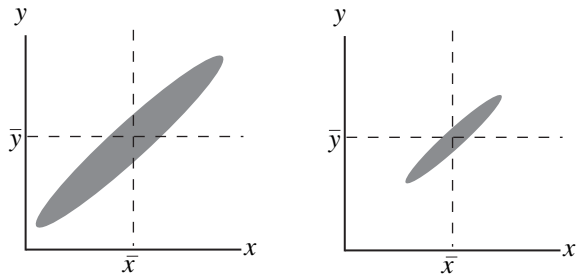

図 4: 同じ相関係数をもつ分布

すなわち

$$L = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2 \tag{2}$$

が最小になるように a と b を決定します（n はデータの組の数です）．このような a と b を求めるには，(2) 式を a と b でそれぞれ偏微分し，それらを両方とも 0 とおいた方程式を解きます．付録 1 のようにして解くと，結果は

$$\begin{aligned} b &= \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x^2}} \\ a &= \bar{y} - b\bar{x} \end{aligned} \tag{3}$$

が得られます．この方法を**最小二乗法**といい，このようにして得られる 1 次式 $y = a + bx$ を y の x 上への**回帰方程式**，あるいは**回帰直線**といいます．また，b は回帰直線の傾きで，これを**回帰係数**といいます．

決定係数

各 x_i に対して，回帰直線上で対応する y の値，すなわち $a + bx_i$ を

$$\hat{y}_i = a + bx_i \tag{4}$$

と表すことにします．このとき，実際のデータにおける y_i と $\hat{y}_i$ の差を残差といい，d_i で表します．残差とは，回帰方程式と x_i の値を使って，y_i の値を $\hat{y}_i$ と予測したとき，予測によって表現できなかった部分を表しています．残差について，r_{xy} を x と y の相関係数とすると

$$\sum d_i^2 = \sum (y_i - \hat{y}_i)^2 = (1 - r_{xy}^2) \sum (y_i - \bar{y})^2 \tag{5}$$

が成り立ちます（導出は付録）．つまり，r_{xy}^2 が 1 に近づくほど y_i と $\hat{y}_i$ の差は小さくなり，$r_{xy}^2 = 1$ のときは残差が 0 となります．すなわち，最小二乗法で求めたモデルによって，y が x から完全に正確に決定されることになります．このことから，r_{xy}^2 を決定係数とよびます．

決定係数の意味は，次のように説明できます．(5) 式を少し変形して

$$1 - r_{xy}^2 = \frac{\sum d_i^2}{\sum (y_i - \bar{y})^2} = \frac{\sum d_i^2 / n}{\sum (y_i - \bar{y})^2 / n} \tag{6}$$

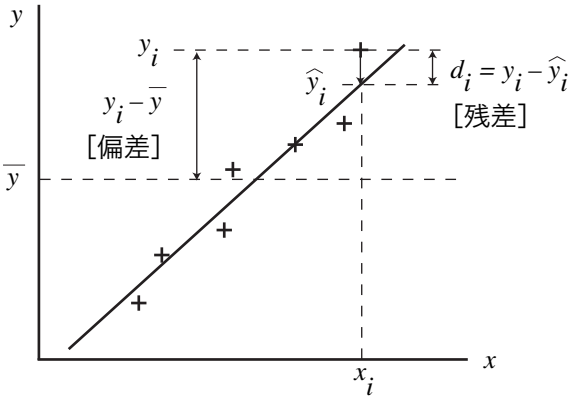

図 5: 偏差と残差

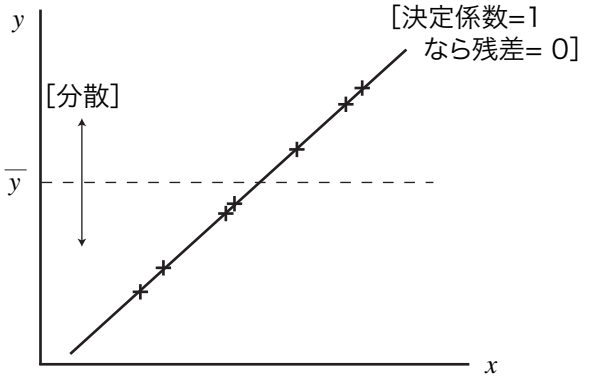

図 6: 決定係数の意味

としてみます．(6) 式の右端の分母は，y 全体の平均からの各 y_i のへだたり，すなわち偏差の 2 乗の平均で，つまり y の分散を表しています．一方，分子は，残差の 2 乗の平均になっています．残差は「線形モデルによる予測結果からの隔たり」ですから，分子は「線形モデルによる予測結果を中心とするばらつき具合」を表しています（図 5）．

したがって，$(1 - r_{xy}^2)$ は「もともとの y のばらつき具合に対する，線形モデルからのばらつき具合の割合」を示す値ということになります．線形単回帰では，「データが散布図上にばらついている」という状況を，「好き勝手にばらついているのではなく，線形モデルで表される直線に沿ってばらついている」と説明しています．しかし，線形モデルで完全に表されたわけではなく，直線から見てもデータはいくらかばらついていますから，上の説明で完全に説明がついているわけではありません．こう考えると，r_{xy}^2 は「直線からのばらつきは，もともとあった y の分散に比べて，何%減少しているのか」を示す値ですから，r_{xy}^2 は「線形単回帰によって，データのばらつきの何%の説明がついたか」を表しています．もし $r_{xy}^2 = 1$ ならば，分散が 100%減少して残差 = 0 ということですから，データのばらつきは線形単回帰によって 100%説明がついた，ということを意味しています．これは，相関係数 = ±1 のときに，散布図上の点が直線上に完全に並んでいることに対応しています（図 6）．

なお，相関係数が 0.7 以上であれば，決定係数はほぼ 0.5 以上になって，回帰直線からのばらつきはもとの分散の半分以下になります．したがって，確かに回帰直線を引く意味がある，すなわち，線形モデルで表すことに意味がある，はっきりとした相関があるといえることになります．ですから，相関係数 0.7 ぐらいが「中くらいの相関」を表します（相関係数 0.5 が中くらいの相関を表すのではありません）．

ところで，y が x によって完全に正確に決定される，つまり決定係数が 1 であるということは，言い方を変えれば「(x_i, y_i) の組になっているデータのうち，x_i さえわかれば，y_i は計算で求められるから，データとして記録する必要がない」ことを意味します．また，決定係数が 1 に近ければ，「x_i がわかれば，y_i の値はだいたい見当がつく」ことになります．このような考え方は，あとの講義で説明する「主成分分析」や「因子分析」で重要な意味を持ちます．

今日の演習

今日の例（緯度 x と気温 y の関係）で，(1) 長野－鹿児島間のデータのみ，(2) 札幌－那覇の全データ，をそれぞれ使って $x, y, \sum x_i^2, \sum x_i y_i$ を計算し，回帰方程式を求めて下さい．(1)(2) の決定係数の違いには，どのような意味があるでしょうか？

付録１：回帰方程式の導出

(2) 式を展開すると（以下，$\sum$ の添字を省略します），

$$L = \sum_{i=1}^{n}\{y_i - (a + bx_i)\}^2 = \sum y_i^2 - 2b\sum x_i y_i - 2a\sum y_i + na^2 + 2ab\sum x_i + b^2\sum x_i^2 \tag{A1}$$

であり，a, b で偏微分してそれぞれ 0 とおくと

$$\begin{aligned}\frac{\partial L}{\partial a} &= -2\sum y_i + 2na + 2b\sum x_i = 0\\ \frac{\partial L}{\partial b} &= -2\sum x_i y_i + 2a\sum x_i + 2b\sum x_i^2 = 0\end{aligned} \tag{A2}$$

となり，それぞれ整理すると，

$$\begin{aligned}na + (\sum x_i)b &= \sum y_i\\ (\sum x_i)a + (\sum x_i^2)b &= \sum x_i y_i\end{aligned} \tag{A3}$$

という連立方程式（正規方程式といいます）が得られます．ここで，x, y それぞれの平均を

$$\bar{x} = \frac{\sum x_i}{n},\ \bar{y} = \frac{\sum y_i}{n} \tag{A4}$$

とおいて代入すると

$$\begin{aligned}na + n\bar{x}b &= n\bar{y}\\ n\bar{x}a + (\sum x_i^2)b &= \sum x_i y_i\end{aligned} \tag{A5}$$

となります．(A5) 式の上段の式から

$$a = \bar{y} - b\bar{x} \tag{A6}$$

が得られます．また，(A5) 式の上段の式を $\bar{x}$ 倍して下段の式から引くと

$$(\sum x_i^2 - n\bar{x}^2)b = \sum x_i y_i - n\overline{xy} \tag{A7}$$

となるので，

$$b = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}^2} \tag{A8}$$

が得られます．

付録２：残差と決定係数の関係の導出

残差の定義から

$$\sum d_i^2 = \sum(y_i - \hat{y}_i)^2 = \sum\{y_i - (bx_i + a)\}^2 \tag{A9}$$

で，さらに (A6) 式を用いると

$$\begin{aligned}\sum d_i^2 &= \sum\{y_i - (bx_i + (\bar{y} - b\bar{x}))\}^2\\ &= \sum[(y_i - \bar{y})^2 - 2b(y_i - \bar{y})(x_i - \bar{x}) + b^2(x_i - \bar{x})^2]\end{aligned} \tag{A10}$$

となります．これに，付録３で説明する

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{A11}$$

を代入すると

$$\begin{aligned}
\sum d_i^2 &= \sum(y_i - \bar{y})^2 - 2\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\sum(y_i - \bar{y})(x_i - \bar{x}) + \left\{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right\}^2 \sum(x_i - \bar{x})^2 \\
&= \sum(y_i - \bar{y})^2 - 2\frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2} + \frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2} \\
&= \sum(y_i - \bar{y})^2 - \frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2} \\
&= \sum(y_i - \bar{y})^2 - \frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}\sum(y_i - \bar{y})^2
\end{aligned} \tag{A12}$$

となります．ここで相関係数の定義を用いると

$$\sum d_i^2 = \sum(y_i - \bar{y})^2 - r_{xy}^2 \sum(y_i - \bar{y})^2 = (1 - r_{xy}^2)\sum(y_i - \bar{y})^2 \tag{A13}$$

が得られます．

付録３：**(A11)** 式の導出

この導出には，

$$n\bar{x} = \sum x_i,\ n\bar{y} = \sum y_i \text{ すなわち } n\overline{xy} = \sum \overline{xy} = \sum x_i\bar{y} = \sum \bar{x}y_i \tag{A14}$$

という関係を用います（x や y は $\sum$（総和）に関して定数であることに注意してください）．

本文 (3) 式の分子は，(A14) 式の関係を用いると

$$\begin{aligned}
\sum x_i y_i - n\overline{xy} &= \sum x_i y_i - \sum \overline{xy} \\
&= \sum x_i y_i - 2\sum \overline{xy} + \sum \overline{xy} \\
&= \sum x_i y_i - \sum x_i\bar{y} - \sum \bar{x}y_i + \sum \overline{xy} \\
&= \sum(x_i - \bar{x})(y_i - \bar{y})
\end{aligned} \tag{A15}$$

となり，また分母も同様の関係を用いて

$$\begin{aligned}
\sum x_i^2 - n\bar{x} &= \sum x_i^2 - \sum \bar{x} \\
&= \sum x_i^2 - 2\sum \bar{x}\cdot\bar{x} + \sum \bar{x}^2 \\
&= \sum x_i^2 - 2\sum x_i\cdot\bar{x} + \sum \bar{x}^2 \\
&= \sum(x_i - \bar{x})^2
\end{aligned} \tag{A16}$$

が得られ，両者から (A11) 式が得られます．