2009 年度前期　データ解析序説　第５回

データの関係を知る (2) ー 回帰と決定係数

多変量データと多変量解析

回帰分析は，２つ以上の変量の組で表されるデータがあるとき，ある変量と他の変量との関係を求める方法です．「関連の強さ」を調べる相関分析と違い，回帰分析では，一方の変量によって他方の変量が決まるという関係があるとき，「ある変量の変化を，もう一方の変量の変化で説明するための関数を求める」という考え方をします．今回は，一番基本的な回帰分析である線形単回帰について説明します．

線形単回帰ー直線のあてはめ

前回の講義で用いた，各都市の緯度と気温のデータ，およびその散布図をもう一度見てみましょう (表 1，図 1)．散布図上のデータは，好き勝手にばらついているわけではありません．前回説明したように，緯度と気温の間には負の相関関係があります．そこで，これらの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「y は x によって説明される」といい，x を説明変数，y を被説明変数といいます．また，この関係を y の x 上への**回帰**といいます．この例の場合，明らかに散布図上で右下がりの直線となるような関係がありそうです．だからといって，散布図上に＋印の列が完全に直線上に並んでいるわけでもありません．では，どういう直線をひけばよいのでしょうか．

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，すなわち

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

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，(1) 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．

> 「a と b それぞれで偏微分する」とは，次のような意味です．微分とは，関数のグラフ上のある点での接線の傾きを求めることです．そこで，(1) 式の L を a, b の２つの変数の関数と考えると，この関数は a, b のどちらについても２次関数で，a^2, b^2 の係数がいずれも正ですから，そのグラフは a, b どちらの軸でみても下に凸の放物線で，すなわち図 2 のような曲面になります．「a と b それぞれで偏微分する」というのは，L を a だけの関数・b だけの関数

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

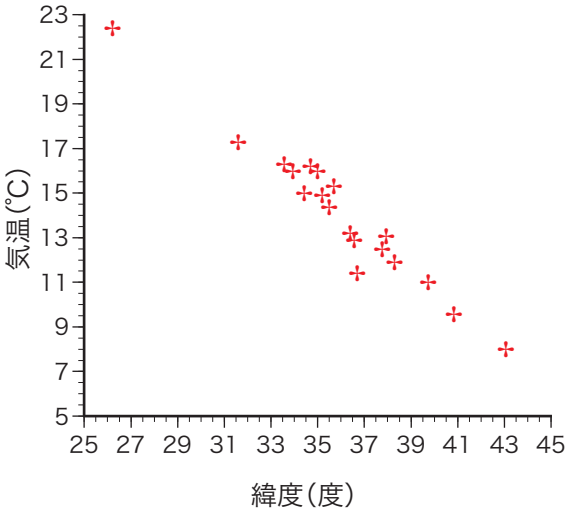


図 1: 散布図：緯度と気温の関係

とみなしてそれぞれ微分することで，曲面上のある点で，a 軸方向の接線の傾き・b 軸方向の接線の傾きを求めることになります．曲面上で，どちらの偏微分も 0 になる点は，曲面の底にしかありません．ですから，どちらの偏微分も 0 になるときの a, b の値が，L を最小にする a, b の値です．

(1) 式を展開すると（以下，$\sum$ の添字を省略します），

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

であり，a, b で偏微分してそれぞれ 0 とおくと

$$\begin{aligned} \frac{\partial L}{\partial a} &= -2 \sum y_i + 2na + 2b \sum x_i = 0 \\ \frac{\partial L}{\partial b} &= -2 \sum x_i y_i + 2a \sum x_i + 2b \sum x_i^2 = 0 \end{aligned} \tag{3}$$

となり，それぞれ整理すると，

$$\begin{aligned} na + (\sum x_i)b &= \sum y_i \\ (\sum x_i)a + (\sum x_i^2)b &= \sum x_i y_i \end{aligned} \tag{4}$$

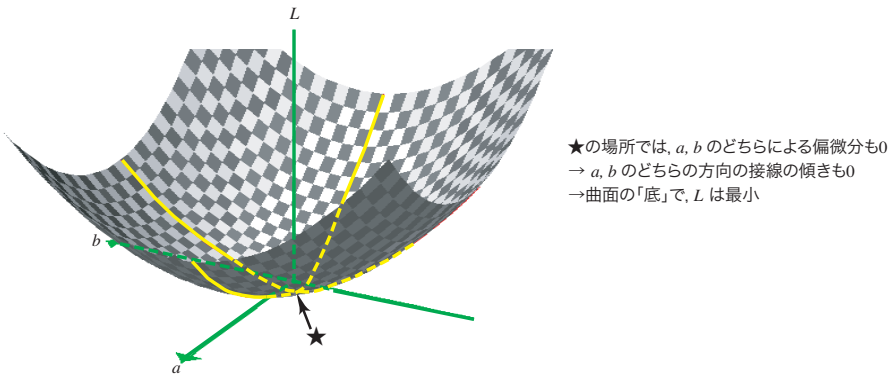


図 2: 偏微分と関数の最小値

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

$$\bar{x} = \frac{\sum x_i}{n},\ \bar{y} = \frac{\sum y_i}{n} \tag{5}$$

とおいて代入すると

$$\begin{aligned} na + n\bar{x}b &= n\bar{y} \\ n\bar{x}a + (\sum x_i^2)b &= \sum x_i y_i \end{aligned} \tag{6}$$

となります．(6) 式の上段の式から

$$a = \bar{y} - b\bar{x} \tag{7}$$

が得られます．また，(6) 式の上段の式を $\bar{x}$ 倍して下段の式から引くと

$$(\sum x_i^2 - n\bar{x}^2)b = \sum x_i y_i - n\overline{xy} \tag{8}$$

となるので，

$$b = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}^2} \tag{9}$$

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．なお，(7) 式を $y = a + bx$ に代入すると

$$y - \bar{y} = b(x - \bar{x}) \tag{10}$$

となりますから，散布図上で回帰直線は「傾きが b で点 $(\bar{x}, \bar{y})$ を通る直線」になります．

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

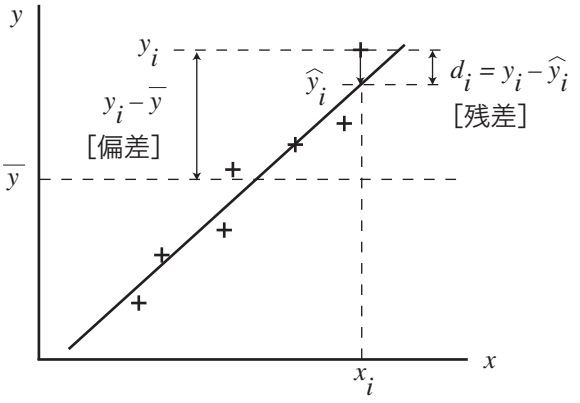


図 3: 偏差と残差

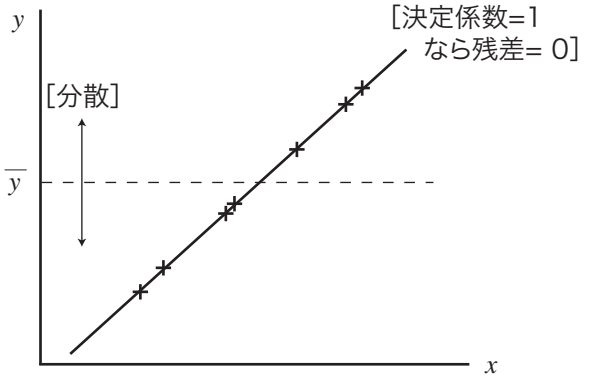


図 4: 決定係数の意味

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（前回の講義参照）とすると

$$\sum d_i^2 = \sum (y_i - \hat{y}_i)^2 = (1 - r_{xy}^2) \sum (y_i - \overline{y})^2 \tag{11}$$

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．

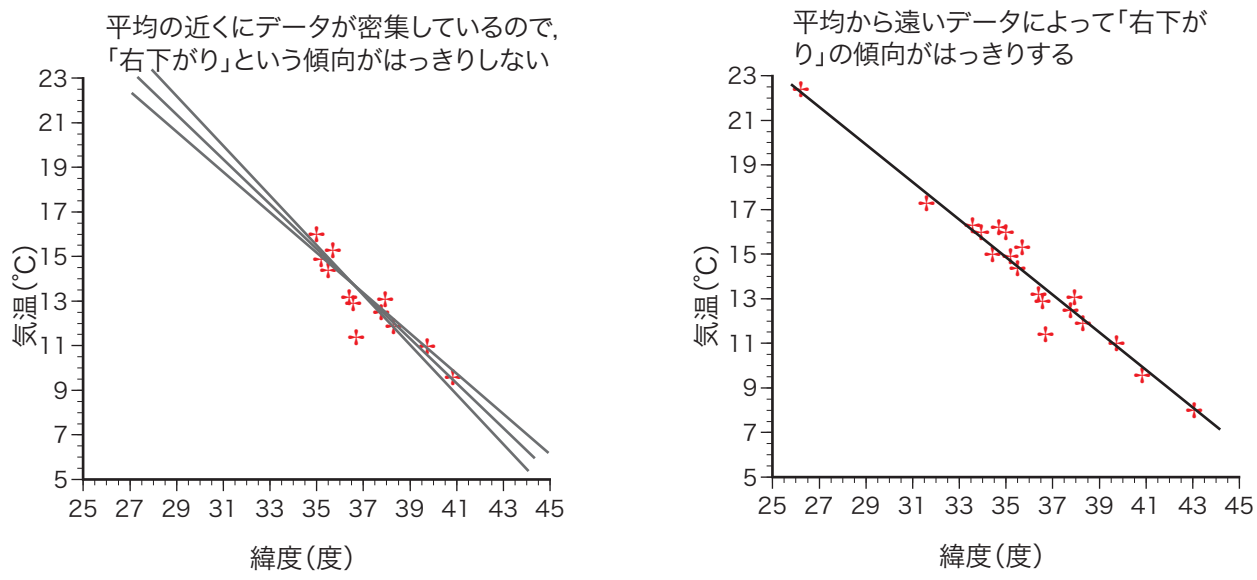
決定係数の意味は，次のように説明できます．(11) 式を少し変形して

$$1 - r_{xy}^2 = \frac{\sum d_i^2}{\sum (y_i - \overline{y})^2} = \frac{\sum d_i^2 / n}{\sum (y_i - \overline{y})^2 / n} \tag{12}$$

としてみます．(12) 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 3）．

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 4）．

前回の講義で，「相関係数 $r_{xy} = 0.5$ は，中程度の相関ではなくほとんど相関が無いことを示す．相関係数 $r_{xy} = 0.7$ であれば，一応相関があるといえる」という説明をしましたが，その根拠はこの決定係数にあります．相関係数 $r_{xy} = 0.5$ のとき，決定係数 $r_{xy}^2 = 0.25$ ですから 25% の減少で，もとの y の分散の 75%は回帰直線からの残差にそのまま残っています．相関係数が 0.7 以上であれば，決定係数はほぼ 0.5



(a) 長野〜鹿児島のデータを使った場合　　(b) 札幌〜那覇のデータを使った場合

図 5: 第 4 回の演習問題の例

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るので，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ほどの，はっきりとした相関があるといえることになります．

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．

前回の演習問題の例

前回の演習問題で，長野〜鹿児島のデータを使った時の相関係数は −0.844，札幌〜那覇のデータを使ったときの相関係数は −0.974 でした．したがって，前者の決定係数は 0.712，後者の決定係数は 0.949 となります．データの大半は共通なのにこのような違いがあるのは，後者のほうには，札幌・那覇という平均から離れたデータがあるために，「散布図のうえで右下がり」という傾向がよりはっきりしていることによります（図 5）．

付録 1 ：残差と決定係数の関係の導出

残差の定義から

$$\sum d_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum \{y_i - (bx_i + a)\}^2 \tag{A1}$$

で，さらに本文 (7) 式を用いると

$$\begin{aligned}\sum d_i^2 &= \sum\{y_i - (bx_i + (\bar{y} - b\bar{x}))\}^2 \\ &= \sum[(y_i - \bar{y})^2 - 2b(y_i - \bar{y})(x_i - \bar{x}) + b^2(x_i - \bar{x})^2] \end{aligned} \tag{A2}$$

となります．これに，付録２で説明する

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{A3}$$

を代入すると

$$\begin{aligned}\sum d_i^2 &= \sum(y_i - \bar{y})^2 - 2\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\sum(y_i - \bar{y})(x_i - \bar{x}) + \left\{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right\}^2 \sum(x_i - \bar{x})^2 \\ &= \sum(y_i - \bar{y})^2 - 2\frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2} + \frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2} \\ &= \sum(y_i - \bar{y})^2 - \frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2} \\ &= \sum(y_i - \bar{y})^2 - \frac{\{\sum(x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum(x_i - \bar{x})^2\sum(y_i - \bar{y})^2}\sum(y_i - \bar{y})^2 \end{aligned} \tag{A4}$$

となります．ここで相関係数の定義を用いると

$$\sum d_i^2 = \sum(y_i - \bar{y})^2 - r_{xy}^2\sum(y_i - \bar{y})^2 = (1 - r_{xy}^2)\sum(y_i - \bar{y})^2 \tag{A5}$$

が得られます．

付録２：(A3) 式の導出

この導出には，

$$n\bar{x} = \sum x_i,\ n\bar{y} = \sum y_i \text{ すなわち } n\overline{xy} = \sum\overline{xy} = \sum x_i\bar{y} = \sum\bar{x}y_i \tag{A6}$$

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 (9) 式の分子は，(A6) 式の関係を用いると

$$\begin{aligned}\sum x_iy_i - n\overline{xy} &= \sum x_iy_i - \sum\overline{xy} \\ &= \sum x_iy_i - 2\sum\overline{xy} + \sum\overline{xy} \\ &= \sum x_iy_i - \sum x_i\bar{y} - \sum\bar{x}y_i + \sum\overline{xy} \\ &= \sum(x_i - \bar{x})(y_i - \bar{y}) \end{aligned} \tag{A7}$$

となり，また分母も同様の関係を用いて

$$\begin{aligned}\sum x_i^2 - n\bar{x}^2 &= \sum x_i^2 - \sum\bar{x}^2 \\ &= \sum x_i^2 - 2\sum\bar{x}\cdot\bar{x} + \sum\bar{x}^2 \\ &= \sum x_i^2 - 2\sum x_i\cdot\bar{x} + \sum\bar{x}^2 \\ &= \sum(x_i - \bar{x})^2 \end{aligned} \tag{A8}$$

が得られ，両者から (A3) 式が得られます．