相関係数に関する若干の考察

二　宮　正　司

０．は　じ　め　に

自然科学や社会科学などあらゆる分野で，所与の標本データにもとづいて，複数個の変量間の関係を調べる必要性が頻繁に生じる。例えば，自動車通行量と交通事故件数，知能指数と試験成績，コレステロール値と血圧，肺癌と喫煙，あるいは心臓病と喫煙・飲酒など，２つの変量の間，または１つの変量と複数の変量の間に関係があるかどうかを検証する必要がしばしば起こってくる。

一般に，量的な増減を共にするような関係は，通常，「相関関係」と呼ばれている。しかし，相関関係は，いろいろな関係の中での特殊なものである。また，２つの変量間に相関が見出されたからと言って，一方の変量が他方の変量に依存して決まるという「因果関係」の存在を必ずしも意味していない。反面，因果関係があれば必ず相関関係は存在する。さらに，２変量の動きがそれぞれが別の原因に依存しているにもかかわらず，相関があるかのような現象が生じる「偽りの相関」に惑わされないように注意する必要もある。

本稿で考察の対象とする相関係数は，ピアソンの積率相関係数である。ノンパラメットリク法である順位相関係数については，ここでは対象としていない。相関の説明は，通常，回帰の説明の中で展開されるか，回帰を説明した後に別立てで展開されることがほとんど

であるが，本稿では最初から相関そのものを考察し，それとの関連で回帰の考察を行う。

先ずは，相関の幾何的解釈を展開する。すなわち，代数的に表現されているピアソンの相関係数の式を，幾何的に分解しその空間的意味を探ることによって，相関に関わる左脳的理解を得ることを目指している。「相関の方向」の幾何的説明はしばしば目にするが，「相関の強さ」に関する幾何的説明はこれまで目に触れたことがない。

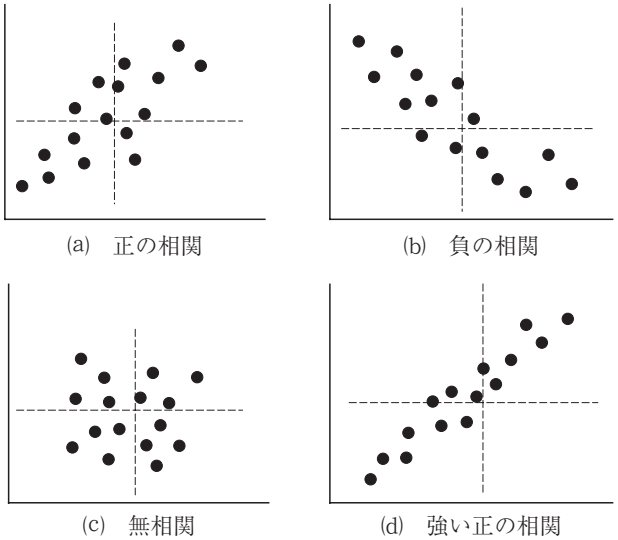
つづいて，この相関係数の計算で基準となっている直線が，最小２乗回帰直線であることを代数的に証明し，２変量の相関係数，回帰の決定係数，回帰の予測値と被説明変数の相関係数，これら３つの係数がまったく一致することを考察する。そして，２変量の場合，説明変数と被説明変数を入れ換えた逆回帰においても，やはり３つの係数が等しくなることを考察する。

次いで，１個の変量と複数個の変量群との間の「重ね合わされた相関」については，もはやピアソンの積率相関係数は適用できないが，多重回帰で推定される決定係数がその役割を果たすことを考察する。その後，相関係数の欠点・問題点や留意すべき点を，数値例を用いて紹介する。最後に，変量が非線形関係にある場合の対処法である線形化変換について，数値例を使いながら説明する。

１．相関関係と単相関係数

２つの量的データが与えられた時，真っ先に行うべき基本的なことは，２次元座標平面上にデータをプロットして散布図を描くことである。図１は，２つの変量のそれぞれの値を x 軸と y 軸に測って，その組み合わせを点で表示したものである。この点の動きによって，２つの変量の定性的関係を探ることができる。図１(a)のように，一方の増加（減少）が他方の増加（減少）を伴う傾向があるとき，定性的に正の相関があるという。逆に(b)の

図１．相関関係の方向と強さ



(a)　正の相関　　(b)　負の相関

(c)　無相関　　(d)　強い正の相関

ように，一方の増加（減少）が他方の減少（増加）を伴う傾向があるとき，定性的に負の相関があると言う。また，(c)のように，２つの変量間になんらかの傾向が見られない場合は，無相関であると言う。相関関係をみる場合，このような「関係の方向」が重要な視点になる。もう一つの視点は「関係の強さ」である。(a)に比べ(d)の方が，２変量間の相関関係は強いと言える。

ところで，相関関係の「関係の方向と強さ」を定量的に示すことができる。２つの変量 x と y の定量的相関の尺度を標本単相関係数，またはピアソンの標本積率相関係数と称し，次のように定義される。ここで，s_x, s_y, はそれぞれ変量 x, y の標準偏差，s_{xy} は２変量の共分散である。

$$r(x,\ y)=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{(n-1)s_xs_y}=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}}=\frac{s_{xy}}{s_xs_y} \tag{1.0}$$

$$s_x=\sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}},\qquad s_y=\sqrt{\frac{\sum(y_i-\overline{y})^2}{n-1}},\qquad s_{xy}=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

このようにして定義される相関係数 r の値の変域は

$$-1\leqq r\leqq 1 \tag{1.1}$$

であり，すべての点が一直線上にあるときのみ−1 または＋1 に等しくなる（数学的証明は後述）。r が正のときは正の相関を，負のときは負の相関を意味し，r の絶対値が１に近づくほど相関の程度は大きいことを意味している。

２．単相関係数の幾何的解釈

いま，２変量 x, y のデータが表１のようであるとすると，その散布図は図２に示される。ちなみに，表１の数値例の相関係数は，0.8064 である。それでは，単相関係数の幾何的解釈を説明するために，まず(1.0)式の分子 $\sum(x_i-\overline{x})(y_i-\overline{y})$ について見てみよう。

分子の $(x_i-\overline{x})$ および $(y_i-\overline{y})$ は，変量 x と y の平均値からの偏差であるから，x 軸と y 軸をそれぞれの平均値分だけ平行移行し，点 $(\overline{x},\ \overline{y})$ を新しい原点にした新しい軸で x, y を測り替えることを意味している。図 3･1 は，x と y の平均値を新しい原点 (13, 8)

表１．２変量データ

x	y
7	6
10	7
10	5
12	9
12	6
13	7
14	10
16	9
17	11
19	10

図２．x, y の散布図

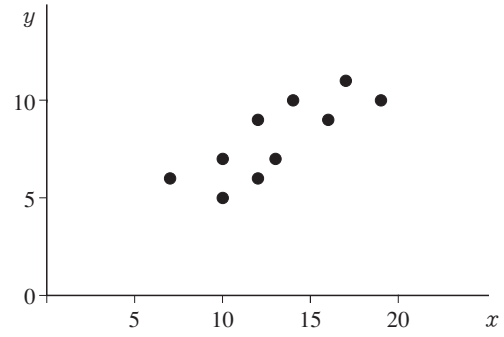


図 3･1．偏差変換された散布図：相関の方向

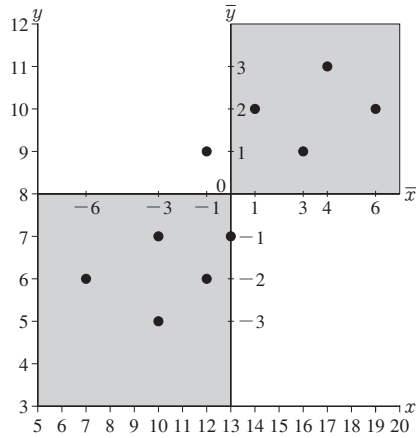
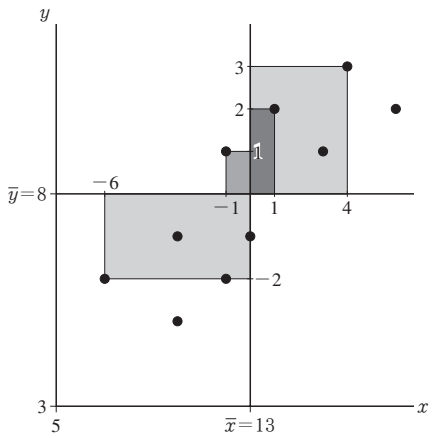


図 3･2．偏差変換された散布図：相関の強さ



にして，図 2 を偏差変換した散布図である。こうして図 3･1 の平面は，第 1 象限から第 4 象限までの 4 つの象限に分割される。そして，偏差変換された新しい変量は，プラスとマイナスの値を持つことになる。

2.1　相関の方向

いま，図 3･1 の網掛に見られるように，変換された変量の新しい x 軸（$x_i-\bar{x}$）と y 軸（$y_i-\bar{y}$）の両軸は，第 1 象限では共に正であり，第 3 象限では共に負である。したがって，点が第 1 象限と第 3 象限（図の網掛け）に散布しているとき，偏差変換された 2 変量の積［交差積］$(x_i-\bar{x})(y_i-\bar{y})$ はいずれも正の値になる。逆に，点が第 2 象限と第 4 象限（図の白抜き）にあるとき，x 軸（$x_i-\bar{x}$）と y 軸（$y_i-\bar{y}$）は一方が正で他方が負に，あるいは一方が負で他方が正になっているから，第 2 象限と第 4 象限にある変量の交差積（$x_i-\bar{x}$）（$y_i-\bar{y}$）はいずれも負の値を持っことになる。

ところで，相関係数の分子は 2 変量の交差積の和 $\Sigma(x_i-\bar{x})(y_i-\bar{y})$ になっている。そこで，多くの点が第 1 象限と第 3 象限に散布して第 2 象限と第 4 象限に散布する点が少なければ，交差積和は全体として正の値になるので 2 変量間に正の相関傾向が見られることになる（図 3･1 および図 1 の(a)と(d))。他方，多くの点が第 2 象限と第 4 象限に散らばっていて 1 象限と第 3 象限に散らばる点が少なければ，交差積和 $\Sigma(x_i-\bar{x})(y_i-\bar{y})$ は全体として負の値を持つので 2 変量間に負の相関傾向が表れる（図 1 (b))。

変換された 2 変量が新しい座標軸の各象限に一様に散布している場合，それぞれの交差積（$x_i-\bar{x}$）（$y_i-\bar{y}$）が正の値も負の値も一様に出現して互いに打ち消し合うために，2 変量の交差積の和 $\Sigma(x_i-\bar{x})(y_i-\bar{y})$ は，ゼロに近い値になるとみられる（図 1 (c))。このとき，2 変量は無相関傾向にあるといえる。

以上のように，交差積の和 $\Sigma(x_i-\bar{x})(y_i-\bar{y})$ に拠って，2 変量の相関関係が正か負かという「相関の方向」を表現できるのである。

2.2　相関の強さ

さらに，交差積和 $\Sigma(x_i-\bar{x})(y_i-\bar{y})$ は，ほんの少しの調整が必要だが，「相関の強さ」の尺度をも包含している。図 3･2 に見られるように，この交差積 $(x_i-\bar{x})(y_i-\bar{y})$ は散布図の各点で描かれる長方形の面積を意味している。この長方形の面積は，散布図の点が新しい原点 $(\bar{x},\ \bar{y})$ からに遠くに散らばっていればいるほど大きくなる。したがって，交差積の和 $\Sigma(x_i-\bar{x})(y_i-\bar{y})$ は各点毎に描かれる長方形の面積の合計を意味している。このために，正の相関傾向が認められる場合では，交差積の和が正の大きいな値をとるならば，原点 $(\bar{x},\ \bar{y})$ からに遠くに散らばった点がより多く存在することを意味しており，散布図は図 1 の(a)よりも(d)のように強い右上がり傾向を形成するわけである。他方，負の相関傾向の場合は交差積が負の値になるが，その交差積の和の絶対値が大きければ大きいほど散布図は強い右下がり傾向になる。

ところが，この交差積の和は，変量・データ $(\bar{x},\ \bar{y})$ の個数とその数値の桁数・単位に大きく依存するという問題を抱えている。1 つ目の問題は，データの数（点の個数）が増加すると分子の交差積和 $\Sigma(x_i-\bar{x})(y_i-\bar{y})$ も比例して大きくなることである。そこで，(1.0)式の標本相関係数は，データの個数に無関係であるようにするために，(1.2)のように，交差積和をデータ数 n（標本データの場合は自由度 $(n-1)$）で割って，交差積の平均値［共分散］になっている。

$$\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{(n-1)}=s_{xy} \tag{1.2}$$

2 つ目の問題は，データの数値の桁数の問題である。例えば，偏差 $(x_i-\bar{x})$ は，数値の桁数が 1 桁のデータの場合ほぼ 1 桁の数値を，5 桁のデータ場合ほぼ 5 桁の値を持つことになる。同様のことが偏差 $(y_i-\bar{y})$ についても言える。したがって，データの桁数が大きいほど交差積 $(x_i-\bar{x})(y_i-\bar{y})$ とその合計が大きくなるというわけである。そこで，この問題を解決するために，x と y の標準偏差 s_x，s_y で割って調整している。すなわち，(1.0)式の相関係数を変形すると以下のように示される。

$$r(x,\ y)=\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_xs_y}=\frac{1}{(n-1)}\Sigma\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \tag{1.0′}$$

各データの平均からの偏差を，そのデータと同じ桁数の標準偏差で割って「標準化」することによって，数値の桁数と単位が異なるデータ・変量から計算する相関係数を「一般的尺度」として定義できるようになる[1)]。

1）この変換は，異なる平均と分散を持つ分布を平均 0，標準偏差 1 の標準分布に変換するための「標準化の公式」と同じ考え方である。

$$z_i=\frac{x_i-\bar{x}}{s_x}$$

3．相関係数の変域

ここでは，まず，相関係数 r の変域範囲 $-1 \leqq r \leqq 1$ を証明する。次に，$|r|=1$ の場合に描かれる直線の特徴について考える。

3.1　相関係数の変域の証明

いま，次のような変数 t の2次関数を考える。

$$f(t)=\sum\{t(x_i-\overline{x})-(y_i-\overline{y})\}^2 \qquad (2.0)$$

この関数は，合計の各項が2乗形式になっているので $f(t) \geqq 0$ である。右辺を展開してこの条件を示すと，(2.1)式になる。

$$f(t)=t^2\sum(x_i-\overline{x})^2-2t\sum(x_i-\overline{x})(y_i-\overline{y})+\sum(y_i-\overline{y})^2 \geqq 0 \qquad (2.1)$$

いま，下記ように定義すると，(2.1)式は(2.2)式ように表示できる。

$$a=\sum(x_i-\overline{x})^2,\ b=-2\sum(x_i-\overline{x})(y_i-\overline{y}),\ c=\sum(y_i-\overline{y})^2$$

$$f(t)=at^2+bt+c \geqq 0 \qquad (2.2)$$

ここで，(2.2)式が成り立つ条件とは，この関数が t 軸と1点で交わる場合か，交点を持たない場合である。すなわち，このような条件は，$b^2-4ac \leqq 0$ であることから，

$$b^2-4ac=4\{\sum(x_i-\overline{x})(y_i-\overline{y})\}^2-4\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2 \leqq 0$$

が成り立つ。これから，下記の(2.3)式が導かれる。

$$\{\sum(x_i-\overline{x})(y_i-\overline{y})\}^2 \leqq \sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2 \qquad (2.3)$$

そして(2.3)式の両辺を右辺で割ると(2.4)式が得られる。

$$\frac{\{\sum(x_i-\overline{x})(y_i-\overline{y})\}^2}{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}=r^2 \leqq 1 \qquad (2.4)$$

(2.4)式の左辺は，相関係数(1.0)式の2乗であるから，$-1 \leqq r \leqq 1$ が成立する。

（証明終わり）

3.2　完全相関の直線

$|r|=1$ のとき完全相関と呼ぶが，散布図の全ての点が直線を描く。その直線はどんな直線であるのだろうか。完全相関の条件は，2次関数(2.2)が0になる（t 軸と接する）場合だから，さかのぼって(2.0)式が $f(t)=0$ である。この式は，合計の各項が2乗形式になっているので，0になるのは下記の(2.5)式を満たす場合のみである。

$$(y_i-\overline{y})=t(x_i-\overline{x}) \qquad (i=1,\ 2,\ \cdots,\ n) \qquad (2.5)$$

すなわち，全てのデータについてこの条件が成り立っている場合に限られる。ここで，相関係数 $r(x,\ y)=1$ として，(1.0)式に(2.5)式を代入して展開すると(2.6)式が得られる。

$$1=r(x,\ y)=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum t^2(x_i-\overline{x_i})^2}}=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{t\sum(x_i-\overline{x})^2} \qquad (2.6)$$

(2.6)式の第1辺と第4辺から，これを t について解くと，(2.7)式が導かれる。

$$t=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}=\frac{s_{xy}}{s_x^2} \qquad (2.7)$$

ここで，t の分母は常に正である。分子は 2 変量の交差積和であるから，「相関の方向」で説明したように，相関が正のときプラスの値，負のときマイナスの値になる。次いで，(2.7)式を(2.5)式に代入すると，偏差変換された原点 $(\overline{x},\ \overline{y})$ を通り勾配が s_{xy}/s_x^2 の直線が求められる。すなわち，

$$(y_i-\overline{y})=\frac{s_{xy}}{s_x^2}(x_i-\overline{x}) \qquad (i=1,\ 2,\ \cdots,\ n) \qquad (2.8)$$

が得られる。この式は，2 変量が完全相関するときの直線を意味している。

ところで，(2.7)式の第 2 辺は，単純回帰推定直線の勾配と同値であることが判明する。すなわち，2 変量が完全相関であるときの直線とは，最小 2 乗回帰直線そのものなのである。(2.7)式の第 2 辺は，次節の単純回帰推定量に関する(3.3)式そのものである。数値例である表 1 と図 2 のデータに対しては，図 4･1 にその直線が図示されており，(2.7)式の t は(3.3)式の b に等しく，その値（勾配）は 0.4576 である。

4．単相関係数と単純回帰推定量

これまでの考察で，相関係数が回帰推定量と深く関わっていることが分かったが，その関わりを更に詳しく検討してみよう。

4.1　単純回帰推定量とその特性

2 変量 $(x_i,\ y_i)$ の線形方程式の推定回帰係数を a，b，および被説明変数 y_i の予測値を $\hat{y}_i$，推定誤差を e_i とすると，単純回帰方程式は(3.1)式および(3.2)式とのように定義される。その最小 2 乗推定量には，(3.3)式から(3.5)式までの特性がある $(i=1,\ 2,\ \cdots,\ n)$。

$$y_i=a+bx_i+e_i \qquad (3.1)$$

$$\hat{y}_i=a+bx_i \qquad (3.2)$$

$$b=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} \qquad (3.3)$$

$$\overline{y}=a+b\overline{x} \qquad (3.4)$$

$$\sum e_i=0,\quad \sum x_ie_i=0,\quad \sum\hat{y}_ie_i=0 \qquad (3.5)^{2)}$$

2）単純回帰の最小 2 乗法を満足する条件は，以下の正規方程式である。

$$\sum y_i=na+b\sum x_i \qquad ①$$

$$\sum x_iy_i=a\sum x_i+b\sum x_i^2 \qquad ②$$

ところで，(3.5)式のそれぞれは，(3.1)式と(3.2)式を代入して，次のように展開できる。

$$\sum e_i=\sum(y_i-a-bx_i)=\sum y_i-na-b\sum x_i \qquad ③$$

$$\sum x_ie_i=\sum x_i(y_i-a-bx_i)=\sum(x_iy_i-ax_i-bx_i^2)=\sum x_iy_i-a\sum x_i-b\sum x_i^2 \qquad ④$$

$$\sum\hat{y}_ie_i=\sum(a+bx_i)e_i=a\sum e_i+b\sum x_ie_i \qquad ⑤$$

ここで，③式は①式を代入して，④式は②式を代入してそれぞれが 0 になる。⑤式の第 1 項は③式により 0，第 2 項は④式より 0 であるから⑤式全体が 0 となる。

したがって，また，(3.2)式と(3.4)式から(3.6)式が，(3.1)式と(3.4)式から(3.7)式が導き出される。

$$\hat{y}_i-\overline{y}=b(x_i-\overline{x}) \tag{3.6}$$

$$y_i-\overline{y}=b(x_i-\overline{x})+e_i \tag{3.7}$$

4.2　単純回帰の決定係数

(3.7)式に(3.6)式を代入すると(3.8)式が導かれる。この式は，(3.1)式を平均からの偏差で分解した形になっている。また，この関係は，(3.9)式で表示されるように，偏差の2乗和についても成り立つ[3)]。

$$y_i-\overline{y}=(\hat{y}_i-\overline{y})+e_i \tag{3.8}$$

$$\Sigma(y_i-\overline{y})^2=\Sigma(\hat{y}_i-\overline{y})^2+\Sigma e_i^2 \tag{3.9}$$

そこで，(3.9)式の左辺で両辺を割ると，(3.9.1)式が得られる。

$$1=\frac{\Sigma(\hat{y}_i-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}+\frac{\Sigma e_i^2}{\Sigma(y_i-\overline{y})^2} \tag{3.9.1}$$

ところで，(3.9)式は次のように説明できる。

（y の総変動）＝（回帰で説明できる変動）＋（回帰で説明できない変動）

ここで，（y の総変動）に対する（回帰で説明できる変動）を $R^2(y|x)$ と定義すると，(3.10)式が成り立つ。ただし，$(y|x)$ は，y の x に対する回帰を意味している。

$$R^2(y|x)=\frac{\Sigma(\hat{y}_i-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}=1-\frac{\Sigma e_i^2}{\Sigma(y_i-\overline{y})^2} \tag{3.10}$$

R^2 は（単）決定係数と称され，回帰推定直線がデータに対してどれ程うまく当てはまっているかを測る尺度となっている。例えば，図4･1は表1と図2のデータの回帰直線を描いており，その推定式と決定係数を図中に示している。決定係数 $R^2=0.6503$ の意味することは，y の総変動（動き）の約65％をこの直線で説明できているということである。そしてまた単純線形回帰では，R^2 は，被説明変数 y が説明変数 x に依存（関係）する程度を測る尺度にもなっている。

また，(3.9.1)式から R^2 が1を超えないことが分かる。また，(3.10)式の第2辺から，その分子も分母も非負であるから，R^2 が非負であることが分かる。すなわち，

$$0\leqq R^2(y|x)\leqq 1 \tag{3.11}$$

が常に成り立つ。

3）(3.9)式の左辺に(3.8)式を代入して展開すると，

$\Sigma(y_i-\overline{y})^2=\Sigma\{(\hat{y}_i-\overline{y})+e_i\}^2=\Sigma\{(\hat{y}_i-\overline{y})^2+e_i^2+2(\hat{y}_i-\overline{y})e_i\}=\Sigma(\hat{y}_i-\overline{y})^2+\Sigma e_i^2+2\Sigma(\hat{y}_i-\overline{y})e_i$

ここで最後の式の第3項は0となることが証明される。すなわち，第3項を次のように展開して，(3.5)式を代入すればよい。

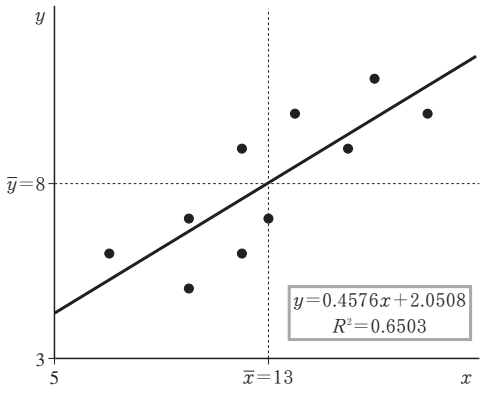
$\Sigma(\hat{y}_i-\overline{y})e_i=\Sigma(\hat{y}_ie_i-\overline{y}e_i)=\Sigma\hat{y}_ie_i-\overline{y}\Sigma e_i=0$

表 2･1．回帰と予測値

説明変数 x	被説明変数 y	予測値 $\hat{y}$
7	6	5.3
10	7	6.6
10	5	6.6
12	9	7.5
12	6	7.5
13	7	8.0
14	10	8.5
16	9	9.4
17	11	9.8
19	10	10.7

$r(x,\ y)=r(\hat{y},\ y)=0.8064$

図 4･1．回帰直線



$\sqrt{R^2(y|x)}=\sqrt{0.6503}=0.8064=r(x,\ y)$

4.3　単決定係数と単相関係数

データが 2 変量の場合，回帰推定される決定係数が，実は，2 変量の相関係数に等しくなる。単決定係数(3.10)式に(3.6)式を代入すると，(3.10.1)式が得られる。

$$R^2(y|x)=\frac{\sum(\hat{y}_i-\overline{y})^2}{\sum(y_i-\overline{y})^2}=\frac{\sum(b(x_i-\overline{x}))^2}{\sum(y_i-\overline{y})^2}=\frac{b^2\sum(x_i-\overline{x})^2}{\sum(y_i-\overline{y})^2} \qquad (3.10.1)$$

さらに，この式に(3.3)式を代入して展開すると，(3.10.2)式が導かれる。

$$R^2(y|x)=\frac{b^2\sum(x_i-\overline{x})^2}{\sum(y_i-\overline{y})^2}=\frac{\left\{\dfrac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})}\right\}^2\sum(x_i-\overline{x})^2}{\sum(y_i-\overline{y})^2}$$

$$=\frac{\{\sum(x_i-\overline{x})(y_i-\overline{y})\}^2}{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}=\left\{\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}}\right\}^2=r^2(x,\ y) \qquad (3.10.2)$$

この式は，y の x への回帰の決定係数 $R^2(y|x)$ が，2 変量 $(x,\ y)$ の相関係数の 2 乗 $r^2(x,\ y)$ に等しいことを意味している。すなわち，次式が成り立つ。

$$R^2(y|x)=r^2(x,\ y) \qquad (3.12)$$

数値例では，図 4･1 に示されているように，単決定係数 $R^2(y|x)=0.6503$ だから，その平方根が相関係数に等しくなる。すなわち，$\sqrt{R^2(y|x)}=\sqrt{0.6503}=0.8064=r(x,\ y)$ が成り立っている。

4.4　被説明変数 y と予測値 $\hat{y}$ の相関係数

ここで，予測値 $\hat{y}$ と被説明変数 y の相関係数 $r(\hat{y},\ y)$ が，説明変数 x と被説明変数 y の相関係数 $r(x,\ y)$ と等しくなることも確認しておこう。相関係数(1.0)式の x を $\hat{y}$ と置き換えて，さらにこれに(3.6)式を代入すると，

$$r(\hat{y}|x,\ y|x)=\frac{\sum(\hat{y}_i-\overline{y})(y_i-\overline{y})}{\sqrt{\sum(\hat{y}_i-\overline{y})^2\sum(y_i-\overline{y})^2}}=\frac{b\sum(x_i-\overline{x})(y_i-\overline{y})}{b\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}}=r(x,\ y) \qquad (3.13)$$

この結果の意味するところは，説明変数 x と被説明変数 y の相関係数を，最小2乗回帰推定による予測値 $\hat{y}$ と被説明変数 y の相関係数で説明できるということである。これは，後述の重相関係数でも述べるように，多変量の相関係数を検討するときの重要なポイントになる。

数値例の表2･1では，予測値予 $\hat{y}$ の列を追加して示している。すなわち，予測値は，x に対する直線上の y の値である。表の予測値 $\hat{y}$ と被説明変数 y の相関係数と，x と y の相関係数とが，共に等しく0.8064となっていることが示されている。

5．逆回帰と相関係数

2変量（x，y）の相関係数では，2変量のいずれが説明変数であり，被説明変数であるかが問われない。その理由について，(3.1)式で取り扱った方程式の説明変数と被説明変数を取り換えた線形回帰方程式（逆回帰）を立て，その回帰推定量を検討してみよう。

5.1　逆回帰推定量とその特性

逆回帰方程式の定義と推定量，および推定量の特性は以下の通りである。

$$x_i = a' + b' y_i + e_i' \tag{3.1′}$$

$$\hat{x}_i = a' + b' y_i \tag{3.2′}$$

$$b' = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (y_i - \overline{y})^2} \tag{3.3′}$$

$$\overline{x} = a' + b' \overline{y} \tag{3.4′}$$

$$\sum e_i' = 0, \quad \sum y_i e_i' = 0, \quad \sum \hat{x}_i e_i' = 0 \tag{3.5′}$$

これから次の式が成立する。

$$(\hat{x}_i - \overline{x}) = b'(y_i - \overline{y}) \tag{3.6′}$$

$$(x_i - \overline{x}) = b'(y_i - \overline{y}) + e_i' \tag{3.7′}$$

そして，被説明変数 x についても(3.9′)式が成り立つから，逆回帰の決定係数 $R^2(x|y)$ として(3.10′)式が定義できる。

$$\sum (x_i - \overline{x})^2 = \sum (\hat{x}_i - \overline{x})^2 + \sum e_i'^2 \tag{3.9′}$$

$$R^2(x|y) = \frac{\sum (\hat{x}_i - \overline{x})^2}{\sum (x_i - \overline{x})^2} = 1 - \frac{\sum e_i'^2}{\sum (x_i - \overline{x})^2} \tag{3.10′}$$

ここで，(3.10′)式に(3.6′)式と(3.3′)式を代入して，(3.10.1)式と同様の式の展開を行うことによって(3.10.2′)式が求められ，したがって(3.12′)式が成り立つ。

$$R^2(x|y) = \frac{\sum (\hat{x}_i - \overline{x})^2}{\sum (x_i - \overline{x})^2} = \frac{b'^2 \sum (y_i - \overline{y})^2}{\sum (x_i - \overline{x})^2} = \frac{\left\{ \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (y_i - \overline{y})^2} \right\}^2 \sum (y_i - \overline{y})^2}{\sum (x_i - \overline{x})^2}$$

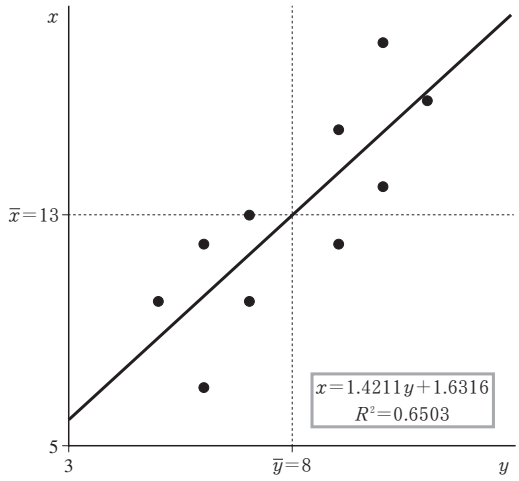
$$= \left\{ \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \right\}^2 = r^2(x, y) \tag{3.10.2′}$$

表 2･2．逆回帰と予測値

説明変数 y	被説明変数 x	予測値 $\hat{x}$
6	7	10.2
7	10	11.6
5	10	8.7
9	12	14.4
6	12	10.2
7	13	11.6
10	14	15.8
9	16	14.4
11	17	17.3
10	19	15.8

$r(y,\ x)=r(\hat{x},\ x)=0.8064$

図 4･2．逆回帰直線



$\sqrt{R^2(x|y)}=\sqrt{0.6503}=0.8064=r(y,\ x)$

$$R^2(x|y)=r^2(x,\ y) \qquad (3.12')$$

表 2･2 の数値例では，y と x の相関係数と，予測値 $\hat{x}$ と被説明変数 x の相関係数とが，共に等しく 0.8064 となっていることが分かる。また，図 4･2 に示されているように，単決定係数 $R^2(x|y)=0.6503$ だから，その平方根が相関係数に等しくなる。すなわち，$\sqrt{R^2(x|y)}=\sqrt{0.6503}=0.8064=r(y,\ x)$ が成り立っている。

5.2　単相関係数と単純回帰推定量

ところで，(3.3)式と(3.3′)式から次の関係式が成り立つ。

$$b\cdot b'=\frac{\{(x_i-\bar{x})(y_i-\bar{y})\}^2}{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}=r^2(x,\ y) \qquad (3.13)$$

(3.13)式はすなわち $\sqrt{b\cdot b'}=|r|$ であるから，単相関係数は推定回帰係数の幾何平均であることがわかる。したがって，線形単純回帰方程式の説明変数・被説明変数が x と y のいずれであるかに関わらず，相関係数はデータに対する直線の当てはめの程度を表すことを意味している。すなわち，(3.12)式と(3.12′)式から(3.14)式が表示できる。

$$r^2(x,\ y)=R^2(y|x)=R^2(x|y) \qquad (3.14)$$

このことは，表 2･1，図 4･1，表 2･2，図 4･2 の数値例で確かめることができる。

6．重決定係数と重相関係数

これまで考察してきた単相関係数は，２変量間の相関の方向と強さを測る尺度であった。他方，１個の変量と複数個の変量との間の関係を測る尺度を考えることができるが，この

場合，ピアソンの積率相関係数に相当するような指標は存在しない。しかし，多重回帰の決定係数の平方根を相関の尺度とすることができる。これを重相関係数と称している。

6.1　多重回帰推定量と重決定係数

いま，被説明変数を y，説明変数を x, z とする線形回帰方程式を考える。この方程式の係数を a，b，c 推定誤差を e とすると，式の定義と推定量の特性は以下の通りになる。

$$y_i=a+bx_i+cz_i+e_i \tag{4.1}$$

$$\hat{y}_i=a+bx_i+cz_i \tag{4.2}$$

$$\overline{y}=a+b\overline{x}+c\overline{z} \tag{4.3}$$

$$\sum e_i=0,\quad \sum x_ie_i=0,\quad \sum z_ie_i=0,\quad \sum \hat{y}_ie_i=0 \tag{4.4}$$

そして，(4.2)式と(4.3)式から(4.5)式が，(4.1)式と(4.3)式から(4.6)式が導かれる。

$$\hat{y}_i-\overline{y}=b(x_i-\overline{x})+c(z_i-\overline{z}) \tag{4.5}$$

$$\hat{y}_i-\overline{y}=b(x_i-\overline{x})+c(z_i-\overline{z})+e_i \tag{4.6}$$

多重回帰の当てはまりの良さを測る尺度としての重決定係数は，単純回帰の場合の(3.9)式，(3.10)式と同様にして，次のように定義される。

$$R^2(y|x\cdot z)=\frac{\sum(\hat{y}_i-\overline{y})^2}{\sum(y_i-\overline{y})^2}=1-\frac{\sum e_i^2}{\sum(y_i-\overline{y})^2} \tag{4.7}$$

6.2　重決定係数と重相関係数

重決定係数は，1個の変量と複数個の変量群との間の「重ね合わされた関係」を測る尺度である。重決定係数の平方根を重相関係数と称している。すなわち，

$$R(y|x\cdot z)=\sqrt{R^2(y|x\cdot z)} \tag{4.8}$$

である。ただし，単相関係数は負の値をとるが，重相関係数はとらない。つまり，

$$0\leqq R\leqq 1 \tag{4.9}$$

となる。その理由は，2変量の間でならば，両方の変化の方向が正であるとか負であるということに意味があるが，1個の被説明変数と複数個の説明変数との間では，変化の方向について解釈することはできないからである。

次に，前述の(3.13)式に関連して予告したが，多重回帰の被説明変数 y とその予測値 $\hat{y}$ の相関係数が，重決定係数に等しくなることを証明しよう。

その前準備として，以下の証明を行う。(4.1)式と(4.2)式から，$y_i=\hat{y}_i+e_i$ だから，

$$\sum(\hat{y}_i-\overline{y})(y_i-\overline{y})=\sum(\hat{y}_i-\overline{y})(\hat{y}_i+e_i-\overline{y})=\sum(\hat{y}_i-\overline{y})^2+\sum(\hat{y}_i-\overline{y})e_i$$

$$=\sum(\hat{y}_i-\overline{y})^2+\sum\hat{y}_ie_i-\overline{y}\sum e_i=\sum(\hat{y}_i-\overline{y})^2+\sum\hat{y}_ie_i-\overline{y}\sum e_i$$

ここで，この式の最終辺の第2項と第3項は0となる。なぜなら，第3項は(4.4)の最初の式から，第2項は(4.4)の最後の式から証明できる[4)]。したがって，以下の(4.10)式が成

4)（(4.4)の最後の式に(4.2)式を代入して展開すると，

$$\sum\hat{y}_ie_i=\sum(a+bx_i+cz_i)e_i=a\sum e_i+b\sum x_ie_i+c\sum z_ie_i$$

この式の最終辺において，各項の何れもが(4.4)式によって0になる。

り立つ。次いで，この結果を予測値 $\hat{y}$ と被説明変数 y の相関係数に代入すると，(4.11)式が導かれる。

$$\Sigma(\hat{y}_i-\overline{y})(y_i-\overline{y})=\Sigma(\hat{y}_i-\overline{y})^2 \tag{4.10}$$

$$r^2(\hat{y}|x\cdot z,\ y|x\cdot z)=\frac{\{\Sigma(\hat{y}_i-\overline{y})(y_i-\overline{y})\}^2}{\Sigma(\hat{y}_i-\overline{y})^2\Sigma(y_i-\overline{y})^2}=\frac{\{\Sigma(\hat{y}_i-\overline{y})^2\}^2}{\Sigma(\hat{y}_i-\overline{y})^2\Sigma(y_i-\overline{y})^2}$$

$$=\frac{\Sigma(\hat{y}_i-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}=R^2(y|x\cdot z) \tag{4.11}$$

以上は，説明変数が2つの場合を説明したが，3つ以上を含む多重回帰一般にも当てはまる。別の言い方をすると，被説明変数 y とその予測値 $\hat{y}$ の重相関係数＝重決定係数は，予測値 $\hat{y}$ の予測精度を示していると言える。

7．相関係数の意味と留意点

相関係数は極めて有効な尺度であるが，利用にあたっては留意すべき点も幾つか持っている。ここでは，相関係数の意味を再考し，その留意点を探ってみよう。

7.1　線形相関

相関係数は，2変量間の「直線的関係」の方向と強さを測る尺度であるにすぎない。この意味で，(1.0)式の相関係数は「線形相関」と呼ばれる。この尺度があらゆる関連を測れる指標ではないことを銘記すべきである。例えば，図5のように関係が非線形である場合，変量間にかなり明確な傾向が見られても，相関係数の数値は，本来持っている値より低くなる。図6に示される散布図に至っては，0に近い値になる。

図7は，放物線 $y=x^2+6x+4$ の x に-3から-9までの値を代入して描いた完全相関の例だが，相関係数を計算すると，完全な0になる。このデータの新しい座標の原点は $(\overline{x},\ \overline{y})=(3,\ 9)$ であるから，$\overline{x}=3$ について左右対称になっている。第1象限と第4象限の対称な点での交差積（面積）は，絶対値は同じだが符号は反対だから，交差積の和はきっちり相殺し合って0になる。第3象限と第4象限についても同様のことが成り立ち，結局，相関係数の分子である交差積（面積）全体の合計が0になる訳である。

図5．非線形散布図

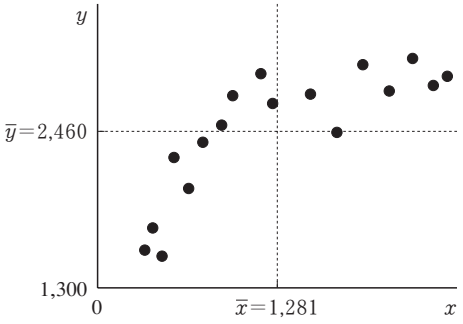


図６．非線形による無相関

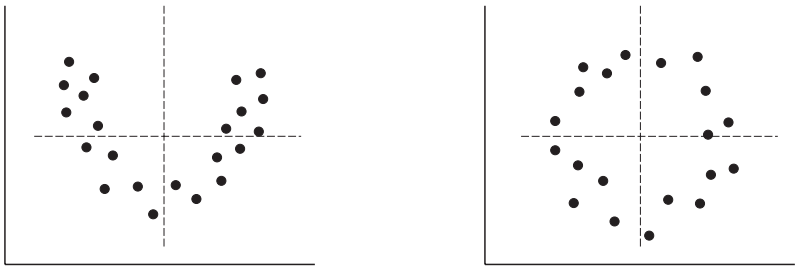
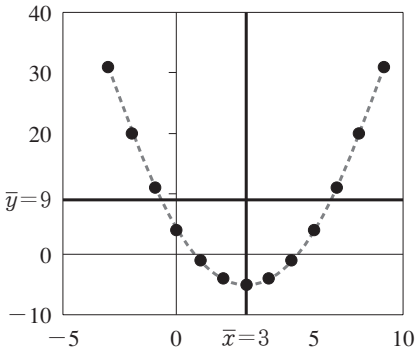


図７．非線形完全相関 $y=x^2-6x+4$



このように，変量間に直線関係が成り立たない場合に，相関係数を使うのは不適切であると言える。また，もし相関係数が低かったとしても，それで相関がないと判断することは危険である。散布図を描いて様子をチェックし，非線形関係がありそうならば，次節で述べるように変量の線形化変換をほどこして，もう一度試みるべきである。初歩的なことであるが，どんな場合でも先ず散布図を描いてみることが不可欠である。

単相関係数の場合と同様に，重相関係数も，１個の変量と複数個の変量群との間の「重ね合わされた線形関係」を測る尺度であって，非線形関係を含む関係一般を測る尺度ではないことに留意すべきである。

7.2　データの同質性

これは統計分析で常に念頭に置くべきことであるが，同質なデータと異質なデータを明確にしておくことが必要である。例えば，図８は，成人72人（男性37，女性35）の身長と体重の散布図である。この図では，男性と女性で明らかに２つのグループに分かれていることが見て取れる。性別の相関係数は男性 0.552，女性 0.537 なのにもかかわらず，このデータ全体の相関係数は 0.809 とかなり高いくなって，過大評価したことになる。また，図 9･1 に示されるように，２つのグループの重なり具合によっては，点の傾向が弱くなり，相関係数は 0.083 と過小評価になる[5)]。このように，同質でない（異質な）データをまとめて計算される相関係数は，間違った結果をもたらすことになる。

図８．身長と体重の散布図

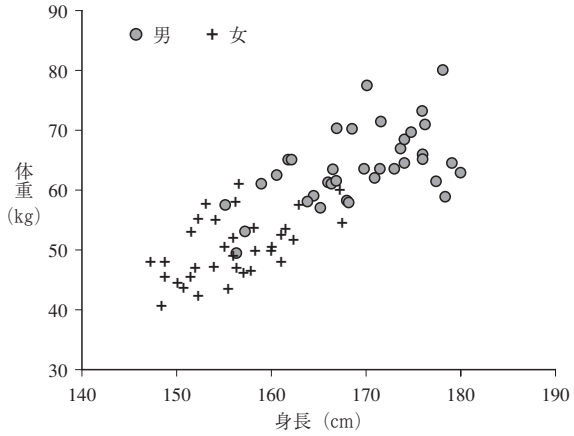


図 9･1．同質ではないデータ（１）

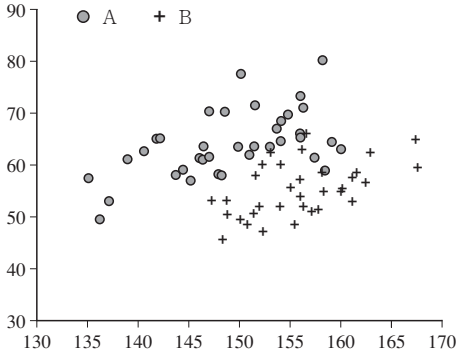
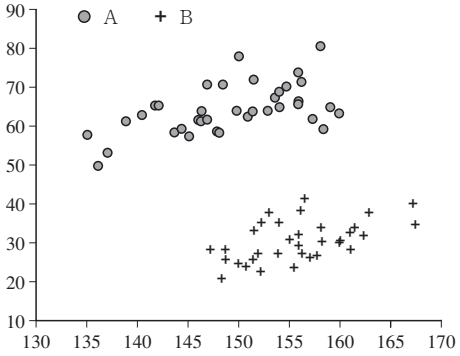


図 9･2．同質ではないデータ（２）



データが同質でない場合のいっそう深刻な問題は，２つの異質なグループが図 9･2 に見られるように散布している場合である[5]。本来ならばＡとＢの各グループが正の相関関係を持っているにもかかわらず，両グループ全体の相関係数は－0.301 と負の値になり，まったく逆方向の相関関係が結果されることになる。

データの異質性をチェックする１つの方法は，万能ではないが，やはり初歩的なことであるが，先ず散布図を描いてみることである。

7.3　外れ値

データが外れ値を持っている場合，それが相関係数に大きく影響を与える。図10･1 と図10･2 は，表１および図２のデータに，それぞれ１つの外れ値を追加して描いたもので

5）この図は，図８の男性グループの相対的位置関係をそのままにして，全体を平行移動している（図のＡ点）。女性グループについても同様の処理をしている（図のＢ点）。したがって，Ａグループ内，Ｂグループ内での相関係数は，それぞれ図８の男性グループ，女性グループと同じである。

図10･1．外れ値（１）

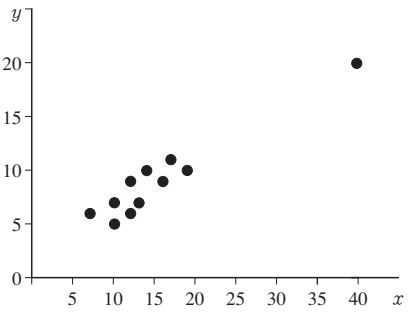
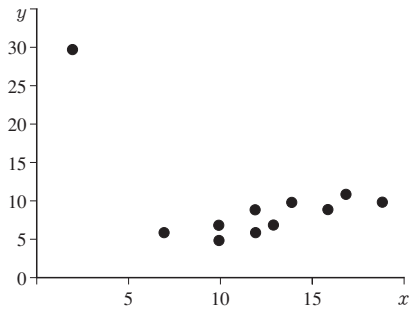


図10･2．外れ値（２）



ある。図２のデータの相関係数は 0.806 であるが，図10･1 では外れ値が１つ追加されることによって，係数は 0.960 に上昇する。他方，図10･2 では，図２にたった１つの外れ値を追加することで相関の方向が逆になり，その係数は－0.503 になる。後者のケースは，計算値のみで相関関係を判断すると，まったく誤った結果を生むことになる。１つの外れ値だけで，正負が入れ替わってしまうわけである。

外れ値は，測定-記録-入力の過程で生じるミスによる場合もあるが，そうでない場合は，その原因をよくチェックする必要がある[6]。外れ値の検出のために先ず試みる方法は，やはり散布図を描いてみることである。

7.4　相関関係と因果関係

相関係数の意味について最も重要なことは，相関係数は単に２変量間の直線的関係の方向と強さを測定する尺度であって，必ずしも因果関係を意味するものではないということである。例えば，夏の季節に気温が上昇するとアイスクリームの売上高が上昇し，また気温上昇に伴って熱射病罹患者数も増加する。ここで，アイスクリーム売上高と日射病罹患者数の相関係数はかなり高くなるだろうが，これで一方が原因で他方が結果であるとは言えない。実は，いずれの変量も気温が原因で，アイスクリーム売上高＝f(気温)，熱射病罹患者数＝f(気温)という関係があるために，結果として両変量が相関しているように見えるのである。これを「偽りの相関」と称する。他の例としては，日本人の血圧と給料との間に強い相関があると言われている。しかし，一般的に年齢に伴って血圧は上昇し，年功序列制が強く残る日本では給料は年齢とともに増加するので，結果として血圧と給料が相関しているように現象するのである。

２つの変量が偽りの相関関係にあるのかどうか，また，因果関係が存在しているのかどうかの判断をする際には，分析対象分野に関する専門知識にも精通して，問題の背後にある知識を十分に身に付けておく必要がある。

6）通常は，外れ値の原因を確かめて，根拠があるならこれをデータから削除する。外れ値が実際にデータとして存在するならば，これに影響されない統計手法（ノンパラメトリック法）を選択する。その代表的な方法としては，スピアマンの順位相関係数，ケンドールの順位相関係数がある。

8．非線形データの相関係数

前節で，変量間に線形関係が成り立たない場合は，線形化変換を試みることを述べたが，ここでは，非線形データの線形化変換について説明する。

表３のデータ（x，y）は，図５を描いた元データである。それを図11･1で再掲している。x と y の平均がそれぞれ 1,281 と 2,460 であり，この図にはさらに回帰直線と回帰曲線が描かれている。また，直線と曲線の推定式と決定係数を，図の右下に示している。散布図の形状を見ても分かることだが，R^2 を比較すると，直線より曲線の方がデータに対する当てはまり具合がいいと判断できる。

図11･1から，２変量の傾向によく当てはまる線は，$y=647.66\times\ln(x)-2053.4$ という曲線（半対数関数）であるらしい。x の自然対数を採って $x=\ln(x)$ とおくことにより線形化変換を試みると，表３の３列目のデータになる。y データはそのままで，x 軸に $\ln(x)$ を採って描いた散布図が図11･2である。線形化変換された変量（$\ln(x)$，y）の散布図が直線傾向を示していることが分かる。そして，この直線の推定係数と決定係数が，図11･1の曲線のそれらと同じになっていることが確認できる。また，表３の２変量（$\ln(x)$，y）の単相関係数を(1.0)式から求めると，(5.0)式になる。そして，それが結局，単決定係数の平方根に等しくなっていることが分かる。

$$r(\ln(x),\ y)=0.8955=\sqrt{0.8019}=\sqrt{R^2(y|\ln(x))} \qquad (5.0)$$

線形化変換について少し述べてみよう。半対数関数 $y=a+b\cdot\log(x)$ の場合，$X=\log(x)$ と変換して，$y=a+b\cdot X$ の線形方程式が得られる。２次関数の場合は，$y=a+b\cdot x+c\cdot x^2$ だから，$Z=x^2$ と変換して，$y=a+b\cdot x+c\cdot Z$ と線形化できる。また，指数関数 $y=a\cdot x^b$ の場合は，両辺の対数をとると $\log(y)=\log(a)+b\cdot\log(x)$ だから，$Y=\log(y)$，$A=\log(a)$，$X=\log(x)$ と変換すると，$Y=A+b\cdot X$ と線形化できる[7)]。

非線形データを線形化する場合注意すべきことは，利用する線形化方程式が数学的に定義されるモデルに限られるということである。したがって，モデル方程式を不適切に適用すると，良い結果が出ないことになる。すなわち，データの傾向にたいして，非線形モデルでどれだけ当てはめができるかに依存しているので，モデルでうまくフォローできる場合に限り，有効であるということである。モデルはあくまでもモデルであって，あらゆるケースに適合するモデルは存在しないので，線形化の限界も認識しておく必要がある。

また，例えば，$y=a+bx+cz$ の多重回帰では，大変に面倒ではあるが，x と y，z と y それぞれの散布図と単相関係数を出すことが重要である。どちらも非線形でないことが判明すれば，そのまま回帰推定し，重決定係数から重相関係数を求めればいい。しかし，被説明変数 y と非線形関係にあると考えられる説明変数があれば，散布図からどのような

7）非線形方程式の代表的なモデルとして，本文で述べた他に分数関数，コブダグラス型関数，ロジスティック関数などがある。

表３．非線形データ

x	y	$\ln(x)$
336	1,575	5.82
395	1,741	5.98
462	1,529	6.14
547	2,262	6.30
649	2,032	6.48
751	2,374	6.62
888	2,500	6.79
965	2,718	6.87
1,166	2,880	7.06
1,250	2,661	7.13
1,522	2,730	7.33
1,711	2,450	7.44
1,900	2,950	7.55
2,087	2,758	7.64
2,254	3,000	7.72
2,400	2,800	7.78
2,500	2,867	7.82

$r(x,\ y)=0.8026=\sqrt{0.6441}=\sqrt{R^2(y|x)}$

$r(\ln(x),\ y)=0.8955=\sqrt{0.8019}=\sqrt{R^2(y|\ln(x))}$

図11･1．非線形データ

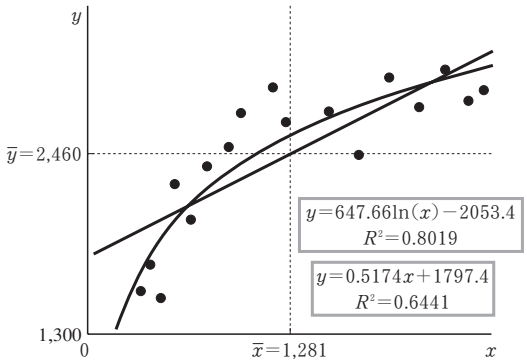
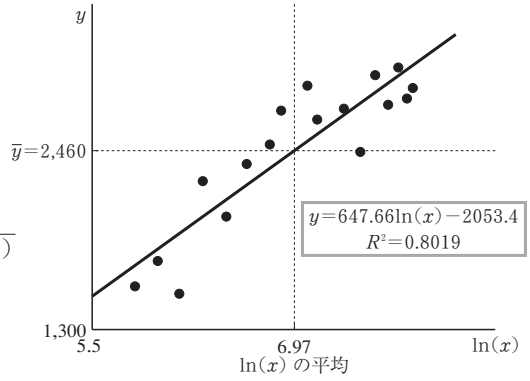


図11･2．線形変換されたデータ



曲線を適用すればいいかを判断し，線形化変換を行うことになる。例えば，$y=a+bx+c\cdot\log(z)$ など，いろいろな形が考えられる。しかしこの場合も，モデルに限りがあるので，どんな場合もうまく線形化変換出来るわけではない。

参　考　文　献

D. B. スーツ（二宮正司他訳）『スーツ統計学』晃洋書房 1979
E. クライツィグ（近藤次郎他訳）『数理統計学２』倍風館 1977
稲葉三男・稲葉敏夫・稲葉和夫『経済・経営統計入門』共立出版 1999
岩田暁一『経済分析のために統計的方法』東洋経済新報社 1983
岩田暁一・木下宗七（編）『テキストブック統計学』有斐閣 1979
刈屋武昭・勝浦正樹『統計学』東洋経済新報社 1994
佐和隆光『初等統計学』新曜社 1976
佐和隆光『数量経済分析の基礎』筑摩書房 1980
白砂堤津耶『初歩からの計量経済学』日本評論社
得津一郎『はじめての統計』有斐閣 2002
二宮正司『統計分析システム』オーム社 1988

宮川公男『基本統計学』有斐閣 1991
森田優三『新統計概論（新訂版）』日本評論社 1999
山本拓『計量経済学』新世社 1995