単回帰と相関

複数の確率変数があり、それらが独立ではなく、何らかの関係を持って変動している時、それらの関係を表す式を作ることを回帰(Regression)と言います。特に確率変数が２つ だけで、１次式でその関係が表せる時には、その回帰を単回帰、直線回帰（linear regression）と言います。具体的な例をあげると、図１に示した散布図をみるとＸと Ｙ の間に何か関係がありそうに見えます。その関係を表す代表的な直線を図の中に引く にはどうしたらよいかということです。

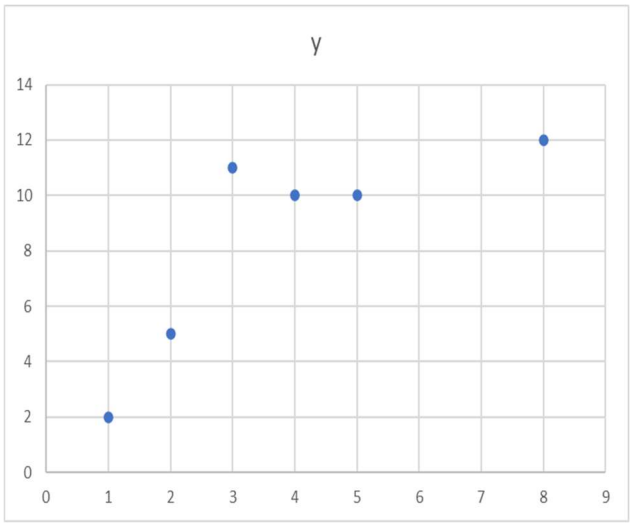


図１．相関がありそうな散布図の例

このデーターを表で表したものが表１です。表の左のカラムには、ＸＹのデーターセッ トの番号を記しました。

表１．相関分析するデータ

No	x	y
1	1	2
2	2	5
3	3	11
4	5	10
5	8	12
6	4	10

分析するのは、図１で、ぼんやりと左から右上がりの直線を中心にしてデーターが散布 しているように見えるということは、どのくらい妥当であるのか。また、直線が引ける とすればその直線の式はどのようになるのかということです。

このことは、すでに学んだ和の分散、差の分散に加えて積の分散（共分散）を論じることに

なるのですが、共分散を単独で論ずるよりは、回帰という具体的な問題を取り扱う方が、かえって共分散という概念を理解しやすいので、単回帰の中で共分散を取り扱います。

仮に仮想的な直線の式を $y = bx + a$ とする。 いつもの例に倣って、

の平均値を、M_x、M_y

xとyの偏差を e_{x_i}、e_{y_i} （i はxとyのデーターセットの番号で $i = 1, \cdots, n$）
と表し、$\bar{y}$を$y = bx + a$を用いてxから予測されるyの値とします

$$\bar{y}_i = b\, x_i + a$$

$$y_i - \bar{y}_i = r_i$$

r_iはxによって説明されない y の残差です。

$$y_i = M_y + e_{y_i}$$

$$x_i = M_x + e_{x_i}$$

ですから、r_iは次のように書けます。

$$r_i = M_y + e_{y_i} - b\left(\left(M_x + e_{x_i}\right) + a\right)$$

変形して

$$r_i = M_y - bM_x - a + e_{y_i} - be_{x_i}$$

平均値では$y = bx + a$が成り立っているとすれば、

$$M_y - bM_x - a = 0$$

ですから

$$r_i = e_{y_i} - be_{x_i}$$

となり、この平方和（残渣平方和）を考えます。

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} \left(e_{y_i} - be_{x_i}\right)^2 = \sum_{i=1}^{n} {e_{y_i}}^2 - 2\,b \sum_{i=1}^{n} e_{x_i} e_{y_i} + b^2 \sum_{i=1}^{n} {e_{x_i}}^2$$

展開した一番右側の式の第 1 項と第 3 項については、すでに和の分散、差の分散の考察を行ってきたそれぞれの変数の平方和です。第 2 項はいままでなじみのないものです。第 2 項の係数を除いた部分$\sum_{i=1}^{n} e_{x_i} e_{y_i}$を自由度で割った値は共分散（covariance）と呼ばれるものです。この項は y が x と関係して変化するために生じた項であり、ためしに、y が x とまったくかかわ りを持たないものとして

$$\sum_{i=1}^{n} e_{x_i} e_{y_i} = 0$$

とすると、

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} {e_{y_i}}^2 + \sum_{i=1}^{n} \left(be_{x_i}\right)^2$$

となり

$$SS_{y+bx} = SS_y + SS_{bx}$$

という、我々が使い慣れた公式になります。

また、y が $y = bx + a$ という式で完全に説明される、言い換えれば、y がすべて $y = bx + a$ y という直線上に集まっているとすれば、

$$e_{y_i} - be_{x_i} = 0$$

ですから、

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} \left(e_{y_i} - be_{x_i}\right)^2 = 0$$

で、確かに残差平方和はなくなります。これらは極めて重要な情報ですが、先を急いで、b の値の最適化に話を集中します。

数学の問題としては

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} {e_{y_i}}^2 - 2\,b \sum_{i=1}^{n} e_{x_i} e_{y_i} + b^2 \sum_{i=1}^{n} {e_{x_i}}^2$$

という式で、$\sum_{i=1}^{n} {r_i}^2$を最小化する、bを求めると問題で、普通は微分して、極値を与えるbを求めるという、解法を使えばよいわけですが、当たり前で、この講義らしくなくてつまらないので、できるだけやさしい解法を使うという趣旨を尊重して、マニアックに、ここでは微分を知ら ない中学生のために、２次関数の最小問題の解法を使うことにします。

$$\sum_{i=1}^{n} e_{x_i} e_{y_i} = SS_{xy}$$

と表すことにします。

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} {e_{y_i}}^2 - 2\,b \sum_{i=1}^{n} e_{x_i} e_{y_i} + b^2 \sum_{i=1}^{n} {e_{x_i}}^2 = SS_y - 2bSS_{xy} + b^2 SS_x$$

bの２次関数らしく書き直して

$$\sum_{i=1}^{n} {r_i}^2 = SS_x b^2 - 2SS_{xy} b + SS_y$$

$$= SS_x \left(b - \frac{SS_{xy}}{SS_x}\right)^2 + SS_y - \frac{{SS_{xy}}^2}{SS_x}$$

以上より、残渣平方和$\sum_{i=1}^{n} {r_i}^2$を最小化するbは

$$\frac{SS_{xy}}{SS_x}$$

その時の残差平方和の値は

$$SS_y - \frac{{SS_{xy}}^2}{SS_x}$$

となります。

bが求まればaも求まるでしょう。これで話は終わりのようですが、統計の解説なのでそれぞ

れの予測値がどのよう確率的な幅をもっているかを検討しておく必要があります。yのもともとの平方和はSS_yですから、次式によって、回帰式によってどのくらい残差 平方和が減ったことになるのか、その貢献度が表せるでしょう。

$$\frac{SS_y - \left(SS_y - \frac{{SS_{xy}}^2}{SS_x}\right)}{SS_y} = \frac{{SS_{xy}}^2}{SS_x SS_y}$$

この値は、回帰によって減った残差平方和の割合で、これも一種の分散比なのですが、これを寄与率（contribution rate）と言い、r^2あるいはR^2という記号で表すことが多いと思います。

$$r^2 = \frac{{SS_{xy}}^2}{SS_x SS_y}$$

個の平方根が相関係数（correlation coefficient）です。

$$r = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

一方、$SS_y - \frac{{SS_{xy}}^2}{SS_x}$も平方和ですから自由度で割れば、残差分散が計算できます。n個の値を持つ２個のデータ群から合成した値なのでこの自由度は $n-2$です。

$${\sigma^2}_y = \frac{SS_y - \frac{{SS_{xy}}^2}{SS_x}}{n-2} = \frac{SS_y\left(1 - \frac{{SS_{xy}}^2}{SS_x SS_y}\right)}{n-2} = \frac{SS_y(1-r^2)}{n-2}$$

この分散は、予測された直線（回帰直線）の周り の値の母集団の２次の積率（バラツキ・広がり方の度合い）です。y の母集団の平均値が 0 である可能性を検討するのであれば、標準誤差は$\frac{\sigma_y}{\sqrt{n}}$、ですから、以下の式で観測値 z を求め、$n-2$の t 分布表の臨界値と比較すればよいでしょう。

$$z = \frac{M_y - 0}{\frac{\sigma_y}{\sqrt{n}}}$$

しかしこれは、実際にあまり意味のある検定ではありません。yの平均値が 0 であるかないかは誰の目にも明らかですし、もしその必要があるとしても、観察されたyの値から平均値を求め、その標準誤差から y の平均値の予測値が 0 を含む可能性について検討すればよいので、わざわざこんな面倒なことはしないでしょう。しかし、これを特別なxの値の時の、yの予測値についての信頼性の検討に用いるならば多少の意味があるかもしれません。たとえば、$x=0$の時の、yの値(y切片)の信頼限界について考えてみます。つまり、$y = bx + a$という式のa の値の予測値の分布範囲について考えるということです。回帰式を作る時に

は、この直線がxとyと x の平均値を通るものとして考えました。つまり、原点をxとy双方の平均値の座標（M_x, M_y）に移動させて、式の傾きのみに着目して、考察を行ってきたのですから、この推定値は y の平均値からbM_x を差し引いた$M_y - bM_x$になることは直感的に分かります（代数的に証明しても良いのですが、とりあえず先を急ぎます。）。y 切片が$M_y - bM_x$で与えられるとすると、y 切片の予測値の積率は y の分散に等しいか ら、たとえば、予測された 1 次式$y = bx + a$が原点を通らない。すなわち$a \neq 0$であることの確率を、

$$z = \frac{M_y - bM_x}{\frac{\sigma_y}{\sqrt{n}}}$$

として、$n-2$の自由度で t 検定したくなるのですが、それは誤りです。
なぜならば、母集団から抽出したデーターによって変動するのはa推定値だけではないからです。回帰直線の傾きbもaとは独立にデーターによって変動します。y切片の値とは、$x = 0$の時の、yの予測値であり、これらは、a、b両方の値の変動によって変動すします。したがってその予測値の真の値のまわりの 2 次の積率は両方を考慮しなければ ならないことになります。そこで、aについての検討をいったん中断してbの予測値の変動について考えます。
bの値の変動について考えることは、aの値の予測値の変動を考えることに比べてはるかに意味があります。そもそも、aの値の変動について予測し、その妥当性について検討するということは、視点を変えれば、母集団の回帰式が本当に 0 を通るのならば、デー ターから作った y 切片の推定値が$M_y - bM_x$となることがあるか、その確率を問うています。その答えとして、母集団の回帰式が 0 を通る時には 95%の確率での$M_y - bM_x$ 値にはならない。という答えが得られたとしてもあまりうれしくはないでしょう。「ある確率でy切片は 0 であり、回帰直線が原点を通る。」と言える方法があるのならまだしも、帰無仮説が否定できなかった場合には、回帰直線が原点を通る可能性については何もいえないし、帰無仮説が否定されたとしてもせいぜい原点を通らないということがいえるだけでそんなことは回帰直線の値から概ね予想がつきます。たぶん、多くの場合、 人々が関心を持つのは「xとyには関連があるのか」ではないでしょうか。知りたいことは$r = 0$ であるかどうかでしょう。これを相関の検定といいます。

$$r = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

であり、$\sqrt{SS_x}$、$\sqrt{SS_y}$ともに 0 でないことは大前提ですから、$r = 0$ の可能性について検討することは$SS_{xy} = 0$の可能性についてその確率を論ずることであり、共分散が 0 である確率の検討でもあります。さらに言えば

$$b = \frac{SS_{xy}}{SS_x}$$

ですから、$b = 0$、つまり傾きが 0 であることの可能性についての検討でもあります。b の予測値がですから。$b = 0$ と$\frac{SS_{xy}}{SS_x}$の差$\frac{SS_{xy}}{SS_x}$を真のb（b の予測値）のまわりの積率で割って、その値 z を t の臨界値と比較すればよいことになります。

私たちに問われていることは、真のbまわりの積率の求めかたです。ここで、分散分析で、標本集団から推定される平均値が母集団の周りにどのように分布するかを考えた経験が役に立つでしょう。私たちはそれを、母集団の 2 次積率の推定値である不偏分散σ^2を個々のデーターに基づく期待値として計算することによって行いました。今回も同様の考え方ができるでしょう。の 1 つのデーターセットから得られる b の値の予測値を

$$b_i = \frac{e_{y_i}}{e_{x_i}}$$

として（xとy平均値を原点として偏差を考えるので、傾きになります。）、この値と全体から得られた

$$b = \frac{SS_{xy}}{SS_x}$$

との差を論じることにします。

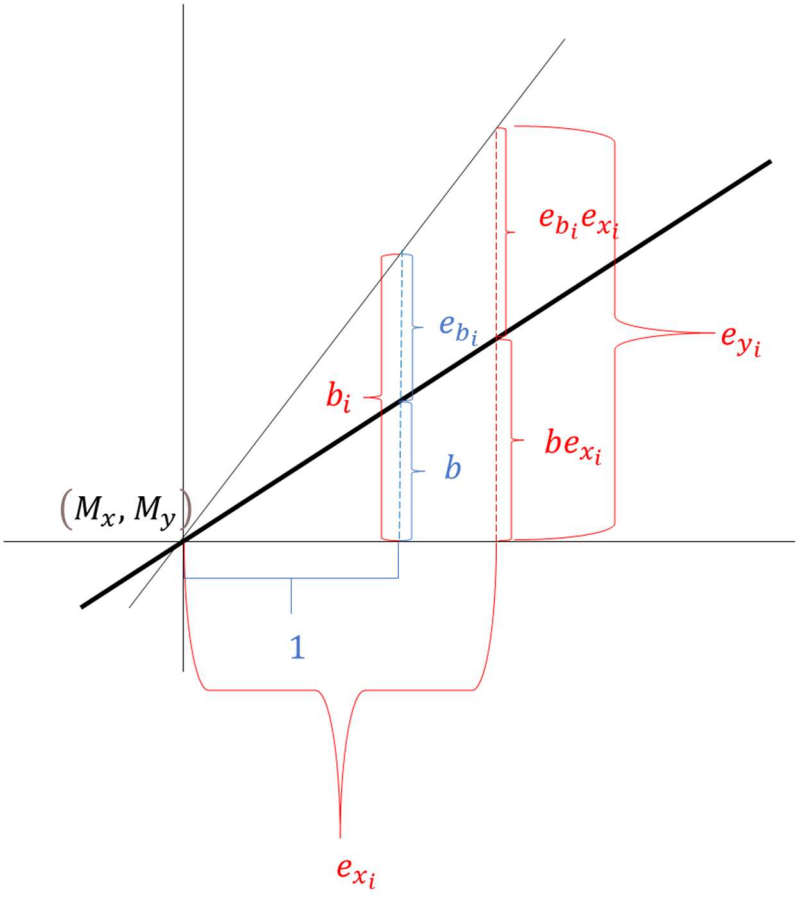


図 1. 傾きの偏差と被説明変数の偏差

回帰式の傾きbとはxが 1 増加した時のyの増加分のことです。また、b_iはe_{y_i}とe_{x_i}の比$\frac{e_{y_i}}{e_{x_i}}$ です。

$$b_i = \frac{e_{y_i}}{e_{x_i}}$$

また、全データから推測した真と思われる傾きは

$$b = \frac{SS_{xy}}{SS_x}$$

で、図 1 に示した通り、$e_{b_i} = b_i - b$ですから、

$$e_{b_i} = b_i - b = \frac{e_{y_i}}{e_{x_i}} - \frac{SS_{xy}}{SS_x}$$

もう少し、変形しておきます。

$$e_{b_i} = \frac{e_{y_i}}{e_{x_i}} - \frac{SS_{xy}}{SS_x} = \frac{1}{e_{x_i}}\left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)$$

$${e_{b_i}}^2 = \frac{1}{{e_{x_i}}^2}\left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2$$

これは、x_iとy_iから予測される傾きb_iと真の傾きbの偏差の 2 乗です。iからnまでのデータ・ペアが選ばれる確率は等しいでしょうから、その総和を求めて、自由度$n-2$で割れば、bの予測値の残差平方和が求まるでしょう。

$$SS_b = \sum_{i=1}^{n} {e_{b_i}}^2$$

なのですが、SS_bを定数と考えて、${e_{x_i}}^2$を移項して、

$$SS_b {e_{x_i}}^2 = \left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2$$

について、左辺右辺でそれぞれ総和を求めた方が何かと安全でしょう。
左辺は

$$\sum_{i=1}^{n} SS_b {e_{x_i}}^2 = SS_b \sum_{i=1}^{n} {e_{x_i}}^2 = SS_b SS_x$$

右辺は

$$\begin{aligned}\sum_{i=1}^{n}\left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2 &= \sum_{i=1}^{n} {e_{y_i}}^2 - 2\sum_{i=1}^{n}\frac{SS_{xy}e_{x_i}e_{y_i}}{SS_x} + \sum_{i=1}^{n}\frac{{SS_{xy}}^2{e_{x_i}}^2}{{SS_x}^2} \\ &= \sum_{i=1}^{n} {e_{y_i}}^2 - 2\frac{SS_{xy}}{SS_x}\sum_{i=1}^{n} e_{x_i}e_{y_i} + \frac{{SS_{xy}}^2}{{SS_x}^2}\sum_{i=1}^{n} {e_{x_i}}^2 \\ &= \sum_{i=1}^{n} {e_{y_i}}^2 - 2\frac{{SS_{xy}}^2}{SS_x} + \frac{{SS_{xy}}^2}{SS_x}\end{aligned}$$

$$= SS_y - \frac{{SS_{xy}}^2}{SS_x}$$

左辺＝右辺で

$$SS_b SS_x = SS_y - \frac{{SS_{xy}}^2}{SS_x}$$

$$SS_b = \frac{SS_y}{SS_x} - \frac{{SS_{xy}}^2}{{SS_x}^2}$$

これが残差平方和で、多くの教科書で箱の形が紹介されていますが、一部の教科書では、さらに変形した形で紹介されています。

$$r = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

$$r^2 = \frac{{SS_{xy}}^2}{SS_x SS_y}$$

ですから、

$$r^2 = \frac{{SS_{xy}}^2}{SS_x SS_y}$$

$$\frac{{SS_{xy}}^2}{SS_x} = r^2 SS_y$$

$$SS_b = \left(SS_y - \frac{{SS_{xy}}^2}{SS_x}\right)\frac{1}{SS_x} = \left(SS_y - r^2 SS_y\right)\frac{1}{SS_x} = (1 - r^2)\frac{SS_y}{SS_x}$$

この方がより洗練されています。

これを自由度$n-2$で割って

$${\sigma^2}_b = \frac{SS_b}{n-2} = \frac{1 - r^2}{n-2}\frac{SS_y}{SS_x}$$

つまり、複雑の計算をした結果、相関による寄与分を差し引いた、分散が、回帰係数bの残差分散だということになります。

$$\sigma_b = \sqrt{\frac{1 - r^2}{n-2}\frac{SS_y}{SS_x}}$$

これは標準誤差になっていますから、

bの推定値$b = \frac{SS_{xy}}{SS_x}$を$\sigma_b = \sqrt{\frac{1-r^2}{n-2}\frac{SS_y}{SS_x}}$ で割って、z値とします。

$$z = \frac{\frac{SS_{xy}}{SS_x}}{\sqrt{\frac{1 - r^2}{n-2}\frac{SS_y}{SS_x}}} = \frac{SS_{xy}}{SS_x}\sqrt{\frac{(n-2)SS_x}{(1-r^2)SS_y}} = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}\sqrt{\frac{n-2}{1-r^2}} = \frac{r}{1-r^2}\sqrt{n-2}$$

$$z = \frac{r}{1-r^2}\sqrt{n-2}$$

となります。これを自由度$n-2$でt検定すれば、$b=0$である確率を論ずることが出来ます。これは、相関の有無を問うていることになるので、それなりに意味がある検定でしょう。

もうどうでも良くなってきましたが、y切片の予測値の母集団のy切片のまわりの２次の積率に話を片付けておきます。傾き bの予測値の２次の積率の議論で示したとおり、xの平均値から遠ざかるにつれて、bの 値の変動の影響は大きくなります。y切片とは、$x = 0$ の時のyの値ですから、平均値か ら平均値（Ｍｘ）分隔たっています。したがって傾きの偏差に由来する偏差は、

$$M_x \frac{\sigma_y}{\sqrt{SS_x}}$$

分散は

$$M_x{}^2 \frac{\sigma^2{}_y}{SS_x}$$

これに、y値の予測値の２次の積率$\frac{\sigma^2{}_y}{\sqrt{n}}$が加わるので、 y切片の予測値の母集団のy切片の値のまわりの積率は

$$\frac{\sigma^2{}_y}{\sqrt{n}} + M_x{}^2 \frac{\sigma^2{}_y}{SS_x} = \sigma^2{}_y \left(\frac{1}{\sqrt{n}} + \frac{M_x{}^2}{SS_x} \right)$$

標準誤差は

$$\sigma_y \sqrt{\frac{1}{\sqrt{n}} + \frac{M_x{}^2}{SS_x}}$$

となります。この値で予測されたy切片の値を割ってＺを求め、自由度 ｎ－２として ｔ 検定を行えばよいでしょう。この方法を用いれば、与えられたxに対するyの誤差範囲も求めることができます。 そのほか、２つの回帰直線の傾きを比較するなど様々な検定が考えられますが、それらも、上記のような方法で、標準誤差を計算したり、あるいは２つの分散を込みにした分散を考えるなどすれば、妥当な検定法を導き出せるはずです。

回帰分析は意外と頼りない

回帰に関してはもっと論じておかなければならないことがあります。

１．飛び離れ値の問題

表 2 に示したｘ、ｙが対になったデーターがあります。これらのデーター間に相関があるかないかを論じます。

表 32, 飛び離れ値のある回帰分析

X	Y
2	1
3	5
5	5
1	3
5	1
20	22

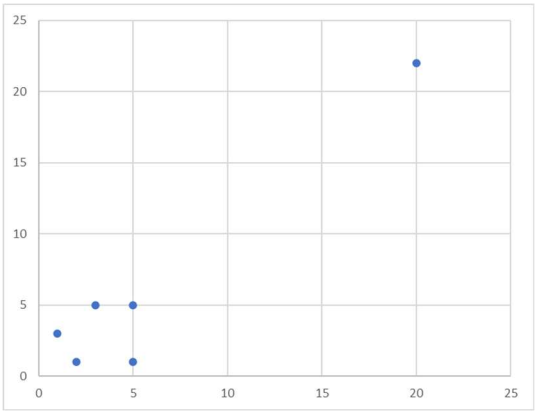


図２．飛び離れ値のあるデータのプロット

回帰分析をする前に当然、グラフを作ってみるでしょう。図 2 にそのグラフを示しま す。確かに相関があるように見えます。実際、相関係数を計算してみると。r=0.956 で、5%以下の危険率で相関は有意になります。たしかに、グラフを見ると、全体としては相関がありそうですが、右上の飛び離れた値を取り除いてみると、他の５つのデーター の間には相関がありそうには見えません。ためしに、この５つのデーターだけで相関分析を行ってみると。r=0.140 でほとんど相関は見られません。右上の飛び離れたデータ ーのために、全体として相関があることになったのです。実際、このような場合、右上 のデーターを取り除いて、解析を行うべきなのか、右上のデーターを加えて解析を行う べきなのか、統計学は教えてはくれません。どうして飛び離れ点が出来たのかを考えなければなりません。それを知ることが出来るのは、統計学ではなくて、研究を行っている当の研究者本人です。 平均値から離れたデーターが大きな影響を持ってしまうのは、この解析で用いているのが最小２乗法による近似を用いているためでもあります。ここでは、わかりやすさを重視して、誤差を最小化する最小２乗法で回帰しました。最近では、確率を最大化する最尤法で近似する方が一般化しているのかもしれません。最尤法には、離れたデーターほど影響が強くなるという問題がありません。しかし、少し計算が複雑になります。

２．相関係数の幾何学的な意味

回帰分析と相関分析とは目的が異なります。回帰分析は、因果関係を持つことがあらかじめ分かっている時に、直線関係を前提に、具体的な関係を示そうとするものです。相関分析が問題にしているのは相関関係があるか否かです。相関関係があっても因果関係があるとは限りません。たとえば、天気が良いと洗濯物が良く乾き、外出する人が多くなります。ですから、洗濯物の乾き方と外出する人の人数には相関関係があります。しかし、洗濯物が乾くから外出する人が多いわけでもないし、外出する人が多いから洗濯物が乾くわけではありません。この場合、相関関係があっても因果関係があるわけではありません。それぞれの分野におけるメカニズムの解明がなければ、因果関係の有無を論ずることはできません。しかし、それでも、数学的にはどちらも相関係数が判断上の重要な指標になっているという意味では共通性がありますので、この項目の最後で、相関係数の幾何学的な意味について考えます。x_i, y_iというペアになったデータがn個在ります。今までの説明では、図３ のように、$x-y$の直交軸で表される平面上の点として それぞれのペアを認識していました。

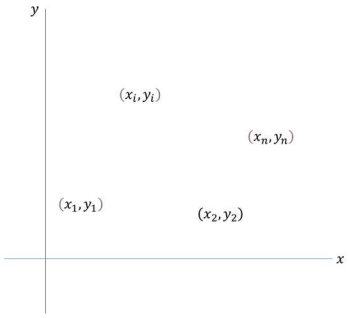


図３． X-Y平面上のデータの散布図

見方を変えると、$\vec{X}=(x_1,\ x_2,\ \cdots,\ x_n)$, $\vec{Y}=(y_1,\ y_2,\ \cdots,\ y_n)$のように成分表示される２つのベクトルと考えることも出来ます。n次元の空間で、互いに直交するn個のベクトルが作る空間で、ϵ_iを単位ベクトルとするベクトル上への$\vec{X}$の写像がx_i、$\vec{Y}$の写像がy_iです。

n次元空間で直交する軸を考えなくてはならないので、図示できませんが、３次元の場合は、図４のように書けます。

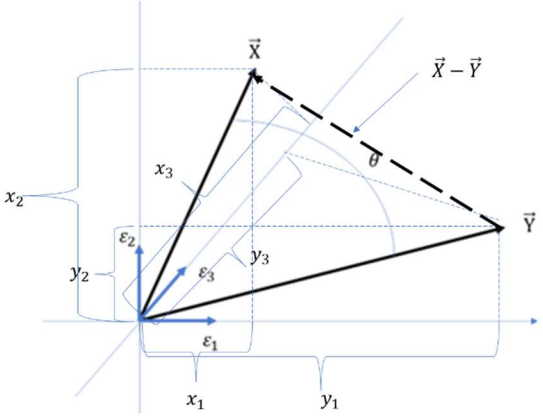
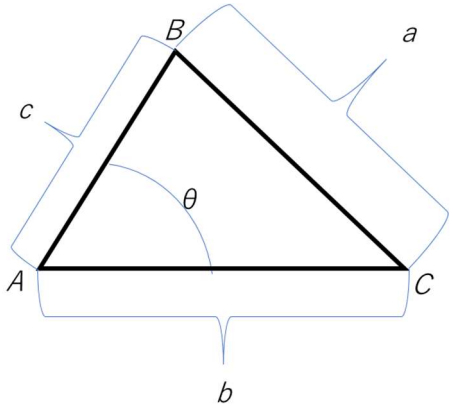


図４. データのベクトル表現

２つのベクトルの内積$\vec{X}\cdot\vec{Y}$の定義は

$$|\vec{X}||\vec{Y}|\cos\theta$$

$|\vec{X}|$はベクトル$\vec{X}$の長さ、θは２つのベクトルのなす角度。



$$a^2 = b^2 + c^2 - 2bc\cos\theta$$

図 4. 余弦定理

個の内積は、図４に示した余弦定理の、右辺第３項にある$bc\cos\theta$に相当するもので

余弦定理をあてはめると、$|\vec{X}-\vec{Y}|^2 = |\vec{X}|^2 + |\vec{Y}|^2 - 2|\vec{X}||\vec{Y}|\cos\theta$

$$|\vec{X}-\vec{Y}|^2 = |\vec{X}|^2 + |\vec{Y}|^2 - 2\vec{X}\cdot\vec{Y}$$

となるので、

$$\vec{X}\cdot\vec{Y} = \frac{|\vec{X}|^2 + |\vec{Y}|^2 - |\vec{X}-\vec{Y}|^2}{2}$$

これを成分表示にすると

$$|\vec{X}| = \sqrt{{x_1}^2 + \cdots + {x_n}^2}$$
$$|\vec{X}|^2 = {x_1}^2 + \cdots + {x_n}^2$$
$$|\vec{Y}| = \sqrt{{y_1}^2 + \cdots + {y_n}^2}$$
$$|\vec{Y}|^2 = {y_1}^2 + \cdots + {y_n}^2$$
$$|\vec{X}-\vec{Y}| = \sqrt{(x_1-y_1)^2 + \cdots + (x_n-y_n)^2}$$
$$= \sqrt{({x_1}^2 + \cdots + {x_n}^2) + ({y_1}^2 + \cdots + {y_n}^2) - 2(x_1y_1 + \cdots + x_ny_n)}$$
$$|\vec{X}-\vec{Y}|^2 = ({x_1}^2 + \cdots + {x_n}^2) + ({y_1}^2 + \cdots + {y_n}^2) - 2(x_1y_1 + \cdots + x_ny_n)$$

なので、

$$\vec{X}\cdot\vec{Y} = \frac{\sum_{i=1}^{n}{x_i}^2 + \sum_{i=1}^{n}{y_i}^2 - (\sum_{i=1}^{n}{x_i}^2 + \sum_{i=1}^{n}{y_i}^2 - 2\sum_{i=1}^{n}x_iy_i)}{2}$$

$$= \sum_{i=1}^{n} x_iy_i$$

$$\vec{X}\cdot\vec{Y}=\sum_{i=1}^{n}x_iy_i=x_1y_1+\cdots+x_ny_n$$

これが、成分表示で書いた、内積の定義式です。二つの定義式は同じものですから、

$$|\vec{X}||\vec{Y}|\cos\theta=\sum_{i=1}^{n}x_iy_i$$

ですが、

$$|\vec{X}|=\sqrt{{x_1}^2+\cdots+{x_n}^2}=\sqrt{SS_x}$$

$$|\vec{Y}|=\sqrt{{y_1}^2+\cdots+{y_n}^2}=\sqrt{SS_y}$$

$$\sum_{i=1}^{n}x_iy_i=SS_{xy}$$

なので、

$$|\vec{X}||\vec{Y}|\cos\theta=\sum_{i=1}^{n}x_iy_i$$

$$\sqrt{SS_x}\sqrt{SS_y}\cos\theta=SS_{xy}$$

$$\cos\theta=\frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

ところで、相関係数rは

$$r=\frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}=\cos\theta$$

ですから、rは２つのベクトルが作る角度θのコサインなのです。ですから、２つのベクトルが重なっていれば

$$\cos 0=1,\ r=1$$

直交していれば、

$$\cos\frac{\pi}{2}=0,\ r=0$$

になります。だから、相関分析というのは、ベクトルの角度を計算しているのです。これの知識をしっかり身に着けておくことは重要です。線形代数的な多変量解析では、ビクトル空間を直交化させるということがしばしば行われます。その時に使われるのが、内積を 0 にする、$\sum_{i=1}^{n}x_iy_i$を 0 にするという計算です（直交化）。

ちなみに、

$$|\vec{X}||\vec{Y}|\cos\theta=\vec{X}\cdot\vec{Y}=\sum_{i=1}^{n}x_iy_i$$

という式は、次の不等式（コーシー・シュワルツの不等式）の証明にもなっています。

$$(\alpha^2+\beta^2+\gamma^2)(\delta^2+\varepsilon^2+\zeta^2)\geq(\alpha\delta+\beta\varepsilon+\gamma\zeta)^2$$

$$\vec{X} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$
$$\vec{Y} = \begin{pmatrix} \delta \\ \varepsilon \\ \zeta \end{pmatrix}$$

とすれば。内積の定義式は、

$$\sqrt{\alpha^2 + \beta^2 + \gamma^2}\sqrt{\delta^2 + \varepsilon^2 + \zeta^2}\cos\theta = \alpha\delta + \beta\varepsilon + \gamma\zeta$$

となって

$$(\alpha^2 + \beta^2 + \gamma^2)(\delta^2 + \varepsilon^2 + \zeta^2)\cos^2\theta = (\alpha\delta + \beta\varepsilon + \gamma\zeta)^2$$

$0 \leq \cos^2\theta \leq 1$だから

$$(\alpha^2 + \beta^2 + \gamma^2)(\delta^2 + \varepsilon^2 + \zeta^2) \geq (\alpha\delta + \beta\varepsilon + \gamma\zeta)^2$$