講 座

使える統計学基礎講座 (第2回)

— 相関と直線回帰 —

光永貴之

独立行政法人農業・食品産業技術総合研究機構中央農業総合研究センター

Lectures of Statistical Logics for Beginners (Part II)

Correlation and Linear Regression

Takayuki MITSUNAGA

National Agricultural Research and Center NARO

1. はじめに

ほとんどの研究者にとって最もなじみの深い統計的手法といえば、相関と直線回帰が挙げられると思います。その理由としては、“二つの変量の関係を見いだす”という命題が研究現場によく見られるためばかりではなく、分散分析等の“何やら厳密な仮説検証型”統計はちょっと難しいという方も相関、回帰と言った“関係の強さを判断してくれる”統計には手法としての親近感や分かりやすさを感じているためであるようにも思います。私の手元に何冊かあります“農作業研究”を開いてみても研究論文、報文24本中10本が相関、回帰を用いています。これは、他の統計手法に比べて際立って多いものです。

しかし、残念なことに相関と回帰のいずれの手法も必ずしも“正しく”用いられているとは限りません。第一回でお話ししました様に、ちょっとした思い違い、勘違いによって本来使うべき状況ではないにも関わらずこれらの手法を用いてしまっている例が見られます。そこで、今回は相関、回帰についてその原理、特徴、注意点などについてお話をしていきます。試験設計、論文作成など

の際の参考にして頂ければ幸いです。

2. ピアソンの積率相関係数

相関係数とは、二つの変量の“関係の強さ”を示す尺度です。言い換えれば、一方の変量の増減が他方の変量の増減にどのくらい（見かけ上）関わっているかを表します。ただし、メカニズムについては考慮しません。従って“相関が高い”ということと“因果関係が強い”ということは全く異なるものです。相関の場合、どちらの変量が主でもう一方が従であると言ったことは特に考える必要がありません。この点が後に説明する回帰との重要な相違点になります。意外かもしれませんが、相関係数と呼ばれるものは何種類も存在します。その中でもっともよく用いられるものが、ピアソン (Pearson) の積率相関係数であり、相関係数と言えば普通はこれを指します。簡単に考え方の基礎を確認しましょう。これは数式を用いて表すと次の様に定義されます。

$$r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)} \sqrt{\text{Var}(y)}} \dots\dots\dots (1)$$

この式の分母は二つの連続変量 x , y それぞれの標準偏差の積です。分子は変量 x , y の共分散を表しています。共分散とは、二つの変量のそれぞれの平均と一つ一つのデータとの差の積を平均したもので、言ってしまうと斜め方向のばらつきの尺度です。積率相関係数は計算が容易であることに加え、値の示していることの分かりやすさも

平成 22 年 2 月 1 日受付

平成 22 年 2 月 13 日受理

Corresponding author

光永貴之 Takayuki MITSUNAGA

〒305-8666 茨城県つくば市観音台 3-1-1

3-3-1 Kannondai, Tsukuba, Ibaraki, 305-8666, Japan

E-mail : aeiou@affrc.go.jp

人気の理由です。 r の値は $-1 \sim 1$ まで変動しますが、符号で関係の性質（一方の大小が他方の大小とどう関係しているか）、数値の絶対値で関係の強さを表します。

さて、今まで二つの変量の“関係の強さ”と書いてきましたが、これは具体的に何を指しているのでしょうか？ 積率相関係数では二変量の散布図を書いた際の“直線らしさ”で評価されます。つまり、相関の高いほど二変量は直線関係に近いということです。このことは逆に、二変量に関係はあるけれども直線関係にはない場合には、積率相関係数を用いてもそれが良質な関係の指標にはならないということを意味します。極端な例では、二変数 x , y が $y = x^2$ の関係にある時に頂点の前後でデータを取れば、積率相関係数は極めて小さくなります。従って、積率相関係数が小さいことと実際に二変量の関係が弱いことは必ずしも一致しないことを理解しておくことは重要です。

積率相関係数を利用する際の注意点としては、はずれ値の影響が大きいことも挙げられます。第一回でもお話ししました様に、データを収集する際にばらつきは避けることができません。データの集団から特に大きく離れている値をはずれ値と呼びます。例として、図1を考えてみましょう。最初にタネを明かしますが、図1のデータは平均値7、分散5の正規乱数（変数 x ）に平均値10、分散5の正規乱数（変数 y ）を割振って積率相関係数を計算しています。二つの正規乱数の平均値を変えた理由はグラフがかっこ良くなるからです。もちろん、変数 x と変数 y には何の関係もありませんから予想される積率相関係数は0です。

図1a) は計算例の1つです。サンプル数は30ですが、積率相関係数は0.1015となり大変小さくほとんどの人は変量間の関係はないと判断すると思います。次に、サンプルの一つを $y = 25$ に置き換えて擬似的なはずれ値を作り、再び積率相関係数を計算してみます（図1b）。すると今度は0.3249となります。多くの人は弱いながらも相関があると判断するでしょう。もちろんこれは極端な例ですが、たった一つの値がぶれるだけで結果が大きく変わる場合があると言うことは理解して頂けるといいます。従って、データのはずれ値を解析に加えるかどうかは、問題としている現象の性質を踏まえて慎重に判断する必要があります。

積率相関係数の安定性をもっと詳しく考えてみ

図1 データにたった一つはずれ値が入ることにより積率相関係数が大きく変わる例。データは変数 x , 変数 y 共に正規乱数を発生させてランダムに割振ったもの。
a) はずれ値がない場合。積率相関係数 $r = 0.1015$ であり、ほとんど相関が見られない。
b) a) のデータから一点だけ取り出し、はずれ値にした場合。相関が a) と比較して大きく高くなっていることが分かる。

ましょう。そのために、先に示した(1)式に戻ってみます。この式は分子が共分散、分母が二つの標準偏差で出来ています。前回のお話を思い出してください。分散やその仲間によって表されているということは、ほとんどの場合正規分布を仮定したパラメトリック手法である（全てではありませんが）ということでした。つまり、積率相関係数には前提があり、それは“二つの変量が共にそれぞれ正規分布に従っている”というものです。従って、二変量が共に正規分布に従っている図2a) のようなデータでは積率相関係数の信頼性は高いのですが、これからデータの分布がずれていくにつれて信頼性は減少していきます。特に、図2b) のような場合には、計算された積率相関係数はその

図2 データの分布と積率相関係数の信頼性の関係.

- 変量 x , 変量 y 共に正規分布に従っている場合. 散布図の中心付近のプロットが最も密度が高く, 周辺に行くにつれて均等に密度が減少している. 積率相関係数は信頼性が高いものが得られる.
- 変量 x , 変量 y 共に正規分布に従っていない場合. 積率相関係数は原点から離れている少数のデータによって大きく影響を受け, 信頼性が高いとは言えない.

値について論ずるには足りません. なぜなら, 分布の集団から外れた少数のはずれ値の影響がとても強く出てしまうからです. 従って, 散布図の疎な部分のデータを大幅に追加して変量 x , 変量 y を共に正規分布に近づける必要があります.

以上述べてきました様に, 積率相関係数は広く用いられている割には, 取り扱いに厄介なところがあります. 実際の利用に当たっては, これらの特質を参考にして慎重に用いることが重要です.

3. 順位相関係数

相関係数とは二つの変量の“関係の強さ”を示す尺度と述べてきましたが, “関係の強さ”を表す指標としては, “直線らしさ”ではなく二変量の変化がどれだけ対応しているか, 言い換えれば“二変量の順位の一貫性”を考えることも出来ます. これに注目した相関係数が“順位相関係数”と呼ばれるもので, 代表的なものとしてはスピアマン

(Spearman) の順位相関係数, あるいはケンドール (Kendall) の順位相関係数が挙げられます. 数式を用いて表すとスピアマンの順位相関係数は次のようになります.

$$\rho = 1 - \frac{6 \sum_{i=1}^N D_i^2}{N(N^2 - 1)} \dots\dots\dots (2)$$

ここで N は二変量のペア数 (サンプル数), D はあるペアでの二変量の順位の差を表します. また, ケンドールの順位相関係数は次の式で定義されます.

$$\tau = \frac{\sum_{i=1}^N P_i - \sum_{i=1}^N Q_i}{N(N-1)} \dots\dots\dots (3)$$

ここで N は二変量のペア数, P はある変量 X の順位 i の際に対応する変量 Y について, それ以上の順位にある X に対応した Y が大きかったものの個数を表します. 一方, Q は Y の順位が i よりも小さかったものの個数です. どちらの順位相関係数も値は $-1 \sim 1$ まで変動しますが, 符号で関係の性質 (一方の大小が他方の大小とどう関係しているか), 数値の絶対値で関係の強さを表します. もしも順位相関係数が 1 だったなら, 二変量の順位は完全に一致していますし, -0.9 であったなら, 一方の変量の順位が上がるほど他方の順位は下がる傾向が大変強いということになります.

積率相関係数が連続変量を取り扱うものであったのに対して, 順位相関係数は順位さえ付けられれば名義変量でも利用可能です. また, 正確な値が不明であったとしても問題ありません. 積率相関係数が変量の正規性を仮定するパラメトリック手法であるのに対して, 順位相関係数は変量の種類だけではなく分布にもとらわれません. 従ってノンパラメトリック手法であると言えます. 例えば, 動物の体サイズと寿命の相関を考えてみましょう. 対象となる動物としてゾウ, イヌ, ハムスター, カブトムシ, ミジンコの5種類を用意します. 積率相関係数を計算するためには, それぞれの体重や寿命のデータを探さなければなりません. しかし, 常識として体サイズ, 寿命ともにゾウ > イヌ > ハムスター > カブトムシ > ミジンコになることは見当がつきます. そこで, 順位相関係数を考えれば体サイズの順位と寿命の順位は完全に一致しているので瞬時に 1 であると結論できま

す。このように変量の種類を問わず、正確なデータである必要が少ないなどの理由から順位相関係数は自然科学よりもむしろ社会科学の分野で広く用いられています。

積率相関係数は直線関係を仮定できない関係では良好な指標となり得ないことをお話しましたが、順位相関係数ではこの問題はある程度解消されます。例として図3を考えてみましょう。ここでは変量 x と変量 y は $y = \log x$ の関係にあります。つまり、変量 y は変量 x によって完全に決められているはずですが、図3a) は変量 x と変量 y をプロットしたものです。積率相関係数は0.9253となり、高いのですが本来の関係を正確に表している訳ではありません。しかし、図3b) にあるように順位に置き換えてプロットすると全くの直線になります。順位相関係数はもちろん1ですから、本来の二変量の関係を正確に表現しているこ

図3 二変量に関係はあるが直線ではない場合の相関係数の比較。変量 x と変量 y は $y = \log x$ の関係にある。

- a) 変量 x , 変量 y のプロットと積率相関係数。相関係数は高いが完全に関係を説明しているとは言えない。
- b) 変量 x , 変量 y を順位に変換してプロットしたものと順位相関係数。関係を完全に説明している。

とになります。もちろん、二次関数等の関係をとる二変量では極値を挟んでしまえば、積率相関係数同様に順位相関係数も適切な指標にはなりません。

順位相関係数もまたはずれ値の影響を受けます。先ほどの図1と同様の操作をしてみましょう。今回も図1と同様に平均値7、分散5の正規乱数(変量 x) に平均値10、分散5の正規乱数(変量 y) を割振っていますが、新たに書き直したものです(図4)。書き直した理由は単に同じ図を使うのを避けたかったからです。図4a) は計算例の1つです。サンプル数は30ですが、積率相関係数は $r = 0.1344$ となり大変小さくなっています。このデータでスピアマン、ケンドールの両順位相関係数を計算すると $\rho = 0.1208$, $\tau = 0.0989$ となり、やはり相関は見られません。次に、サンプル

図4 データにたった一つはずれ値が入ることにより順位相関係数が変わる例。データは変量 x , 変量 y 共に正規乱数を発生させてランダムに割振ったもの。

- a) はずれ値がない場合。積率相関係数 $r = 0.1344$ であり、ほとんど相関が見られない。スピアマン、ケンドールの順位相関係数は $\rho = 0.1208$, $\tau = 0.0989$ となり、同様に相関は見られないと判断できる。
- b) a) のデータから一点だけ取り出し、はずれ値にした場合。全ての相関係数が a) と比較して高くなっているが振れ幅は順位相関係数の方が小さいことが分かる。

の一つを $y=25$ に置き換えて擬似的なはずれ値を作り、再び全ての相関係数を計算してみます（図 4b）。すると今度は $r=0.3170$, $\rho=0.2525$, $\tau=0.1908$ となります。積率相関係数と比べて振れ幅は小さいですが、順位相関係数もやはりはずれ値の影響を受けることが分かります。

以上見てきました様に、総じて順位相関係数は積率相関係数に比べて使い勝手の良い指標であると言えます。しかし、自然科学の分野で順位相関係数を用いている例はあまり多くありません。なぜ順位相関係数が採用されないのかということには、個々の研究シーンでそれなりの理由があるのかもしれませんが、私には研究者の心情として“直線信仰”と言えるものがある様に思えます。つまり、“二変量の関係が密な場合には直線関係に従ってもらいたいし、直線関係こそベストな関係であろう”と言ったスタンスです。このようなスタンスを強く持ちすぎしまうと、現実（＝データ）から現象を読み取るという本来の自然科学のあり方ではなく、予想した現象にデータを当てはめることになってしまい大変危険なことであると言えます。あるいは直線関係であれば、片方の変数から他方の推測も容易であり、実用上も便利であると考えられるのかもしれません。しかし、変量の推測は相関のカバーするところではありません。相関の意味するものはあくまでも二変量間の（見かけの）関係の強さであり、それ以上のものではありません。変量の予測は回帰の行うべきことです。実際の研究現場では、変量の正規性が満たされない場合や関係の有無のみを議論したい場合が意外に多いものです。そのような場合には順位相関係数を使うと良いでしょう。

4. 偽相関と偏相関係数

相関係数は関係の強さを示す尺度であると述べてきましたが、関係の前提となる因果関係を問題にしています。もちろん、相関係数が高いということは多くの場合、何らかのメカニズムが隠されていると判断できるのですが、必ずしも直接関係がない場合でも高い相関が見られることがあります。こうした現象を偽相関と呼びます。

例として図 5 を考えてみましょう。このデータは一人の日本人が一年に食べる牛肉の量とマグロの量の積率相関を見たものです。ただしこれはあくまでも仮想データです。このデータから積率相

図 5 偽相関の例.

- a) 日本人が一年間に食べる牛肉とマグロの量の散布図（仮想データ）。見かけ上の相関は極めて高い。
- b) 牛肉、マグロの消費量と可処分所得の相関を示した図。どちらも相関が高いため見かけ上 a) の高い相関が見られる可能性が否定できない。

関係数を計算すると $r=0.8500$ となり、大変高い相関を示します（図 5a）。この結果から、“マグロ好きは牛肉好き”と直ちに判断できるでしょうか？もちろん、どちらも脂っこいですからその可能性はあります。しかし、この説明に直ちに納得はできません。私たちの生活を考えてみても食べたいものを躊躇なく食べるには先立つものが必要になります。そこで、牛肉消費量、マグロ消費量のそれぞれをその人の可処分所得と相関を取り直します。その結果、それぞれの相関はともに高いことが分かりました（図 5b）。つまり、牛肉、マグロのそれぞれの消費量は可処分所得と関係が高く、かつ同方向に相関するために見かけ上大変高い相関を示すことが予想されます。

本当に“マグロ好きは牛肉好き”は否定されたのでしょうか？これを判断するためには偏相関係数を用いるのが便利です。これは3つの変量（変量1＝牛肉、変量2＝マグロ、変量3＝可処分所得としましょう）があった時に、変量3の影響を取り除いた後の変量1-変量2間の相関係数です。偏

相関係数は次の式で計算されます。

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \dots\dots\dots(4)$$

この式に従って偏相関係数を計算すると $r_{12} = 0.8500$, $r_{13} = 0.8100$, $r_{23} = 0.8450$ ですから 0.5281 となり、元の値に比べると小さいですが、やはりかなりの相関があることが分かります。従って、この場合はある程度は“マグロ好きは牛肉好き”と言っても良さそうです。このように、相関係数の取り扱いには、ある程度の背後関係を抑えておくことが重要です。

5. 相関係数の検定と適切な利用

積率、順位を問わず、相関係数を計算した後に忘れてはならないのは算出された値に意味があるかどうかの確認、すなわち検定です。統計的な表現をするならば、“計算された相関係数は0である”という帰無仮説を棄却できるかどうかの確認です。ここで帰無仮説を棄却できなければ、計算された数値はたまたま出てもおかしくないと判断されて意味のないものになってしまいます。幸いなことに、相関係数の検定は相関係数表を用いれば簡単に行うことができます(表1、表2)。もちろん、実際には相関係数の計算を行う際には何らかの統計ソフトを使う場合がほとんどでしょうから、ソフトが自動的に行ってくれます。参考として、表1に積率相関係数、表2にスピアマンの順位相関係数の検定表を示しました。ここで示されている臨界値は絶対値ですので計算された相関係数の符号は気にする必要がありません。ケンドールの順位相関係数については省略させていただきます。

さて、この二つの検定表からすぐにわかることは相関係数の有意性検定はサンプル数によって完全に決定されるということです。例として、図1の二つの積率相関係数を検定してみましょう。図1a)では $r=0.1015$ 、図1b)では $r=0.3249$ です。サンプル数は両方とも30です。表1からサンプル数30の場合の臨界値を読み取ると有意水準5%では0.361になっています。図1a)、b)の両者ともこの臨界値を下回っているので統計的には0と区別できないことになり、“相関があるとは言えない”と結論されます。先にも述べました様に図1b)の場合では弱いながらも相関があると言いた

表1 ピアソンの積率相関係数の相関係数表

サンプル数	5%水準	1%水準
3	0.997	1.000
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.755	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
15	0.514	0.641
20	0.444	0.561
25	0.396	0.505
30	0.361	0.463
40	0.312	0.403
50	0.279	0.361
60	0.254	0.330
70	0.235	0.306
80	0.220	0.286
90	0.207	0.270
100	0.197	0.256

表2 スピアマンの順位相関係数の相関係数表

サンプル数	5%水準	1%水準
3		
4		
5	1.000	
6	0.886	1.000
7	0.786	0.929
8	0.738	0.881
9	0.700	0.833
10	0.648	0.794
15	0.521	0.654
20	0.447	0.570
25	0.398	0.511
30	0.362	0.467

くなるのですが、サンプル数が30程度ではこのくらいの値では不十分であるわけです。逆に、サンプル数が100の時には、5%水準の臨界値は0.197ですから、かなり弱い相関でも“相関がないとは言えない”と言うことが出来ます。従って、相関の有無を議論したい場合には十分なサンプル数を用意することが重要になります。もちろん、“相関があると判断された”ということと“関係が強い”ということは別の問題です。

次に、二つの積率相関係数の大きさの違いを検定したい場合にはどうするのでしょうか?この場

合には残念ながら多少数式をいじってやる必要があります。最初に二つの積率相関係数を便宜的に r_1, r_2 としてそれぞれ次の式で変換します（フィッシャーの Z 変換と呼びます）。

$$Z_i = \frac{1}{2} \ln \left(\frac{1+r_i}{1-r_i} \right), \quad i=1, 2 \dots\dots\dots(5)$$

この値から次の統計量を計算します。

$$Z = \frac{|Z_1 - Z_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \dots\dots\dots(6)$$

ここで、 n_1, n_2 はそれぞれサンプル数です。最後に (6) 式で得られた数値を自由度 ∞ の t 分布表（ほとんどの統計入門書に載っていますし、インターネットでも見つけることが出来ます）と見比べて有意水準での臨界値より大きければ、有意差あり（二つの相関係数は同じであるとは言えない）と判断します。一般には両側検定を用いるので、見るべき有意水準は 0.025 以下の部分になります。この方法はスピアマンの順位相関係数の検定にも使えますが、ケンドールの順位相関係数では式が変形しますのでこのままでは使えません。藤井（1983）などを参照してください。

実際には、自然科学において二つの積率相関係数（順位相関係数でも）を比較すべき状況と言うのはそれほど多くありません。理由は簡単で、“どちらがより直線らしいか” という問題はいつでもいい場合が多いからです。例として、次のような学会発表があったとしたら皆さんはどのように評価しますか？ウサギの体重と一日に食べる草の量の関係について調べました。この時点でウサギに興味がないと言う答えはここでは NG にしましょう。雌雄によって平均体重が違っているのでオスとメスを分けてプロットしてみました（図 6）。その結果、雌雄いずれも体重と食べた草の重さには有意な相関が見られました。計算された相関係数からはメスの方がちょっと高い相関の様にも見えます。そこで次にオスとメスでどちらが相関が高いのかを比較しました。結論としてメスの方がちょっと相関が高かったです。おしまい。もしもこの発表を大日本ウサギ学会（そんなものはありませんが）で発表したとしたら議論百出になること間違いありません。ほとんどの人は図 6 を見た瞬間にオスの方が急な傾きになっていることに気づくと思い

図 6 ウサギの体重と一日に食べる草の量の散布図（仮想データ）。雌雄で体重が違うので二つに分けた。どちらも有意な相関が見られる。

a) オスの場合、b) メスの場合。

ます。そして、次に思うことは“なぜ？”ということ。あるいは実用的なことを考えれば、飼養するにはメスの方が餌代がかからなくて良いなどと推論するかもしれませんが、いずれにしても、直線らしさの雌雄間での違いについて興味を持つ人は少ないでしょう。

最初に述べました様に、相関は現象の生じるメカニズムについては言及しません。従って、これらの疑問について有益な情報を与えてはくれません。その意味で相関係数は研究の出口を導くものではなく、入り口を示すものであると認識しておくことは重要なことです。もちろん、相関が高いということは二つの変量間の関係が強いということであり、何らかのメカニズムの存在を“予測”することはできます。もっと言ってしまえば、メカニズムの存在を予測するためのツールです。しかし、メカニズムの存在やその原理等を“証明”することはできません。つまり、相関係数の計算

の先には原則として操作実験等のメカニズムを証明する作業が必ず存在するべきです。例えば、図6の場合、新たに出てきた疑問である“ウサギにおいてオスとメスの体重－餌量間の傾きに違いがあるのか？”という疑問には回帰分析を行って傾きの違いを検定する必要があります。あるいは“オスとメスで食べる餌量に違いがあるのか？（同時に体重－餌量の反応に雌雄で違いが見られるのか？）”という疑問には分散分析（共分散分析）を行うことになります。どちらにしろ、これらの検定を行うためには、これらの検定の前提を満たす様に実験デザインをきちんと設定する必要があります。そしてこれらの検定で有意に違うと言えた場合に始めて現象を引き起こす何らかのメカニズムの存在を証明できたことになります。

最後に相関係数を用いた場合の発表のコツについて紹介しておきます。お話ししてきました様に、相関係数は主として研究の入り口を示す尺度です。ですから、学会発表や成績発表、論文等では、データの相関係数について紹介する場合にはできるだけ始めのうちに出示してしまうことが大切です。逆に結論近くに持っていくほど発表に締まりがなくなってしまいます。是非一度お試し下さい。

6. 直線回帰の特徴

二つの変量があった時、片方の変量（説明変数）の大きさによってもう片方の変量（目的変数）の大きさを説明できるような数式を回帰式と呼びます。つまり、回帰を行うのは、原則として目的変数を推定する目的がある場合です。従って回帰は相関と違い、変量間の主従関係がはっきりと決められる必要があります。回帰式は関数であれば原則として何でも構わないのですが、通常用いられている式は一次関数です。従ってこれを特に強調する場合、直線回帰（線形回帰）と呼びます。逆に、二次関数を用いれば二次回帰、ロジスティック関数を用いればロジスティック回帰などと呼びます。

変数間の主従関係は定義式からも明瞭です。直線回帰は以下の式で定義されます。

$$Y - Y_M = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - X_M) \dots\dots\dots(7)$$

説明変数 X, 目的変数 Y に対して X_M , Y_M はそれ

ぞれの平均値を表しています。係数項は分子が二変数間の共分散、分母が説明変数の分散になっています。このことから、回帰直線は相関と違い（式(1)を参照してください）、二変数を等価に扱っている訳ではないことが分かります。もちろん、積率相関係数同様、パラメトリック手法です。従って変数が割合等の正規分布を仮定できない場合にはそのままでは使えません。割合のデータの場合、最近ではロジスティック回帰が主流ですが、もしも直線回帰にこだわるなら、変数変換をする必要があります。変数変換には様々なものがありますが、割合のデータについては、逆正弦平方根変換（逆正弦変換、アークサイン変換とも言います）が一般に使われます。これは一つ一つの割合のデータ t を次の式で変換したものです。

$$\text{Arcsin}\sqrt{t} \dots\dots\dots(8)$$

この変換を行って正規分布により近づけた後に直線回帰を行うのが望ましいと言えます。なぜこの変換を行う必要があるのか、これを行うとどのくらい正規分布に近づくのか、と言った疑問は当然持たれると思いますが、ここでは触れません。藤井（1983）等を参照してください。

回帰直線は手計算で算出できますし、エクセル等のアプリケーションでも簡単に95%信頼限界を含めて作成することができます。また、JMPなどの統計ソフトではクリック操作だけで逆推定（得られた回帰直線を用いて目的変数から説明変数を推定すること）も信頼区間付きで行えます。そういった点で大変お手軽なので多くの研究で当たり前のように行われています。しかし、厳しいことを言えば多くの研究では十分な仮定を満たしていません。その理由は係数項の取り扱いにあります。係数項の分母は説明変数の分散になっていますが、説明変数（説明するための変数ですからコントロールできる前提です）に測定誤差が存在する場合、分散は“本来の”値よりも大きくなってしまいます。すると、分母が大きくなりますので回帰直線の傾きは“本来の”ものよりも小さく推定されます。もちろん全く測定誤差を除去することは不可能ですが、直線回帰を行う場合には、説明変数を厳しくコントロールする必要があることが分かります。その一方で、第一回で説明しました様に、測定誤差は正しく評価されなければなりません。直線回帰の場合、知りたいのは目的変

図7 理想的な直線回帰のデータ。厳密にコントロールされた説明変数に対してばらつきを判断するための複数の目的変数のデータが存在する。

数の値なので、こちらのばらつきは十分吟味できる様に繰り返し測定する必要があります。この二つのことを同時に満たす理想的な直線回帰のデータの取り方は図7の様になります。目的変数のばらつきは回帰直線の信頼限界を正確に判断するために重要になります。もしも、1説明変数あたり一つの目的変数の値しか存在しない場合（相関のときの散布図のようなデータ）ですと、誤差の判断が難しくあまり精度の良い推定ができません。加えて、相関でも見てきましたはずれ値の影響を受けやすくなります。野外データではここまでコントロールするのは難しいかもしれませんが、これに近づける努力は必要です。まして、室内の操作実験では厳密に考えるべきです。

以上述べてきました様に、直線回帰を精度よく行うためには実験設定を厳密に行う必要があります。もしかすると、“そこまで確からしい直線は求めているない”と思う人もいるかもしれませんが、その場合には、積率相関が分かれば十分かもしれません。直線回帰は得られた式によって目的変数を推定するためのもので、推定が必要ない場合にあって行う必要はありません。相関と回帰は自分の目的に応じて使い分けることをお勧めします。

回帰直線が求められると、次の作業として得られた回帰直線に意味が見いだせるかどうかの検定を行う必要があります。回帰直線の検定は、

- (1) 傾きの検定（得られた直線の傾きが0と有意に違うかどうか）
- (2) 切片の検定（得られた切片が0と有意に違うかどうか）

の二種類があります。この二つの検定をクリアした状態で始めて予測式として利用できる回帰直線となります。さらに、2本の回帰直線の傾きが違うかどうかの検定も考えられます。しかし、これらの検定については次回にお話しさせていただきます。

7. 第二回のまとめ

1) 広く用いられているピアソンの積率相関係数は二変量の関係の強さを“直線らしさ”で評価する尺度である。二変量が直線関係にない場合には適切な関係の指標にならない場合があることに注意することが重要である。また、はずれ値の影響を大きく受けるのでデータの取り扱いに慎重になる必要がある。特に、二変量が正規分布をしない場合、少数の集団から離れたデータによって相関係数は大きく影響を受けてしまうのでこうした場合には適切な指標となり得ない。

2) 順位相関係数は二変量の関係の強さを“順位の一致性”で評価する尺度である。変量は順位さえつけられれば連続量である必要はなく、データ分布も気にしなくてよい。はずれ値の影響は受けるが積率相関係数に比べてその大きさは小さくなる。あえて二変量が直線関係にあると言う必要がない場合には使い勝手の良い尺度である。

3) 直接因果関係のない場合でも共通の変数が存在する場合、見かけ上高い相関関係が見られることがある。これを偽相関と呼ぶ。偽相関は偏相関係数を用いることによって共通変数の影響を取り除くことができる。

4) 実際に相関係数を計算した際には最初に計算された相関係数が0と区別ができるかどうかについての検定をかける必要がある。この検定は相関係数表によって簡単に行うことができる。二つの相関係数の違いについての有意性検定はフィッシャーのZ変換を行った後に所定の計算式で統計量を算出して行う。得られた統計量は自由度 ∞ のt検定表によって有意性を判断する。

相関係数は研究の入り口を見つけるのに適した尺度であるが出口を導くものではない。従って、原則としてその後に操作実験等で因果関係を証明する必要がある。

5) 回帰はある説明変数によって、関数関係にある目的変数を推定する手法である。特に一次式を仮定した場合、直線回帰と呼ぶ。相関と違い、

両変数の主従関係ははっきりしている。直線回帰はパラメトリック手法なので正規分布を仮定できないデータ（割合等）には使えない。変数変換（特に逆正弦平方根変換）を行うか、ロジスティック回帰等の手法を用いる必要がある。データを取る際には説明変数を厳しくコントロールして、かつ1つの説明変数に対して複数の目的変数を取ることが望ましい。

今回は、相関と回帰について使用上の注意点を中心にお話しました。広く普及しているこれらの

手法も実際には様々な制約があることを理解して頂ければ幸いです。重要な点は、“制約が強いので使用を控える”のではなく、“使用出来る様に試験設定を整える”と言うことです。次回は直線回帰の続きと分散分析についてお話します。

8. 引用文献

藤井宏一（訳）（1983）：生物統計学（原本：Sokal, R.R. and F.J. Rohlf (1961) Introduction to biometry), 共立出版, 東京, pp.447.