Argonauta 13: 3-23 (2007)

重回帰、偏相関、パス解析

大垣俊一

生態学の論文によく現れる統計処理のうち、多変量解析系のものとしては、主成分分析やクラスター分析のような要素分類的なものと、重回帰など予測、説明的手法がある。前者については以前この Newsletter で論じたので（大垣 1999)、今回はもう一つの、重回帰系の多変量解析を取り上げたい。

重回帰分析とは簡単にいえば、何らかの目的とする変数を複数の要因に関連づける手法である。たとえば漁獲量に対する水域の水温、塩分の寄与を評価する。一方偏相関は、変数どうしが独立でない（たとえば水温と塩分に相関がある）ときに、そうした相互作用を除去して、特定の２変数（水温と漁獲量など）の関係を分析するといった場合に用いられる。

本論ではまず、重回帰、偏相関の基礎として、多くの読者になじみのある単回帰、単相関をおさらいする。その後に重回帰、偏相関の理論を解説し、それらの発展形態としてのパス解析にも言及したい。なお文中で取り上げる具体例は、これまで通り海岸生態学の分野から選んだ。

出発点

重回帰系に限らず多変量解析の出発点は、次のような変数と事例（ケース）の行列表である。

		変数			
		X1	X2	X3	・・
	1	- -	- -	- -	
ケ	2	- -	- -	- -	
｜	3	- -	- -	- -	
ス	：	：	：	：	

表１．多変量解析のデータ表

主成分分析やクラスター分析の場合、たとえば変数 X が調査地点、ケースが種、表中に各地点での各種の個体数が入ると、種相に基づく地点の分類を行うことができる。重回帰では、変数のうち一つ（Y とする）を、他の変数群の関数として構成しようとするので、表は次ページのようになる。ここで X と Y に対しては、次のように様々な呼び名がある。

X： 説明変数（explanatory variable） Y： 目的変数（criterion variable）
　　独立変数（independent variable） 　　従属変数（dependent variable）
　　予測変数（predictor variable） 　　基準変数（criterion variable）

		目的変数	説明変数		
		Y	X1	X2	・・
	1	- -	- -	- -	
ケ	2	- -	- -	- -	
｜	3	- -	- -	- -	
ス	：	：	：	：	

表２．重回帰分析のデータ表

表現の客観性という点では独立－従属がよいのかもしれないが、ここでは変数選択の際に意味が明確でまちがいにくいという実用的な理由で、説明－目的の用語を用いる。言うまでもなくこれらの表現は、「目的」とする要因Ｙをある変数Ｘによって「説明」するという状況を含意している。

単回帰と単相関

表２で、１つの目的変数Ｙに対し、説明変数Ｘも１つしかないという最も単純な場合が単回帰、単相関の適用となる。説明変数が複数ある一般形が今回テーマとする重回帰、偏相関だが、それらの理論では多次元空間をイメージしなければならないなど、しばしば理解がむずかしい。そうした場合でも、単変量の理論を延長することによってある種の納得が得られることがあり、単回帰、単相関を振り返ることには意味があるだろう。

１．単回帰

ある説明変数に対して目的変数の値を平面座標系に打点して両者の関係を見たとき、点が直線に近い形に配置していることがある。このとき、これらの点の間に一本の直線を通してその傾向を最もよく代表させようとするのが回帰という操作である。そのためには、その直線が仮に引けたとして、各点からその直線に対し、y 軸に平行に引いた線分の長さの２乗の和が最小になるようにする（最小２乗法, 図１）。このとき、「y 軸に平行」であって、χ軸に平行でも、直線に垂線をおろすのでもないことに注意する必要がある。このことは、後に述べる「回帰における説明変数と目的変数の非対称性」に反映する。

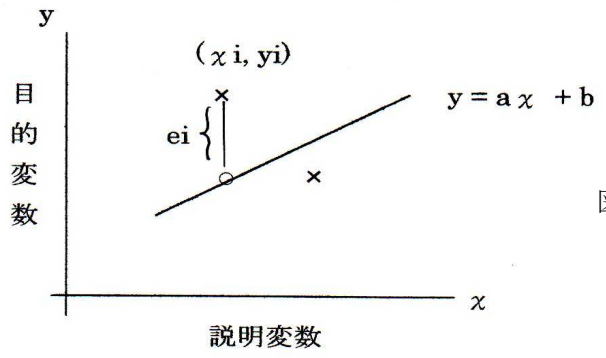

図１．単回帰と最小２乗法

具体的な計算方法としては、図１のように求める直線を　$y = a\chi + b$　とし、各点から y 軸に平行にこの直線に引いた線分の長さを e とすれば、i 番目の点では

$yi = a\chi i + b + ei$ であるから、

$ei = yi-(a\chi i + b)$ となり、ei の全点についての２乗和

$\Sigma ei^2 = \Sigma \{yi-(a\chi i + b)\}^2$ … ①

が最小になるように a、b を決めることになる。このときの ei を残差または誤差、Σei^2 を残差平方和と呼ぶ。「残差」は直線回帰によって説明しきれない残りの部分、「誤差」は直線によって推定したときの、実際の値からのズレというニュアンスがある。

①は a、b についての２次式だから、それぞれについて整理すると、詳細は省くが、

$\Sigma ei^2 = (\ )a^2 + (\ )a + (\ ) = (\ )b^2 + (\ )b + (\ )$

のように書くことができる。a、b が共に変化する状況下で、Σei^2 を最小にするには、上の a の２次式における a と、b の２次式における b、それぞれについて微分した式＝0 とすればよい。２次関数では、微分＝０の点が最小値（グラフの頂点）を示すからである。その結果、

$(\ )a + (\ ) = 0$　　$(\ )b + (\ ) = 0$

の２つの式が得られる。実際には a の式の（　）内に b 、b の式の（　）内に a が入っており、

$a\Sigma\chi i + nb - \Sigma yi = 0$　　$a\Sigma\chi i^2 + b\Sigma\chi i - \Sigma\chi i\, yi = 0$　（n は点の総数）

のようになる。

未知数２つに対し式２つなので、これで a、b は解けるが、これに、回帰直線は全点の平均の位置（χ_m, y_m）を通るという条件（$\Sigma\chi i = n\chi_m, \Sigma yi = n y_m$；$\chi_m, y_m$ は $\chi i, yi$ の平均値）を加えて式を簡単にすると、最終的に次のような回帰式が得られる。

$Y = \{(\Sigma\chi i\, yi - n\chi_m y_m)/(\Sigma\chi i^2 - \chi_m^2)\}X + (y_m - \Sigma\chi i\, yi - n\chi_m y_m)/(\Sigma\chi i^2 - \chi_m^2)\chi_m$

この式の傾き、切片の部分は、すべてもとのデータ表から計算できるから、これで与えられたすべての点から回帰直線が引かれたことになる。

また、a、b を、χ、y の分散、共分散を用いてあらわすと、

χ の分散：　$Sxx = 1/n(\chi i - \chi_m)^2$　、　y の分散：　$Syy = 1/n\Sigma(yi - y_m)^2$

共分散：　$Sxy = 1/n\Sigma(\chi i - \chi_m)(yi - y_m)$　として、

$a = Sxy / Sxx$、　$b = y_m - (Sxy / Sxx)\chi_m$　と書くこともできる。

さらに先の式①に求めた a、b を代入することによって、残差平方和 Σei^2 の最小値 S_0 が、以下のように決まる。

$S_0 = n\, Syy^2 (1 - Sxy^2 / Sxx\, Syy)$

この式の中の、$Sxy^2 / Sxx\, Syy$ は、のちに述べる相関係数、

$r = \Sigma(\chi i - \chi_m)(yi - y_m) / \{\sqrt{\Sigma(\chi i - \chi_m)^2}\sqrt{\Sigma(yi - y_m)^2}\}$

の２乗である。つまり相関係数 r =1 のとき、最小残差平方和 $S_0 = 0$ となり、すべての点が回帰直線に乗る。相関係数については、次項で示すような共分散の意味からの定義もあるが、このように、回帰理論から導くこともできる。

ここで $Sxy^2 / Sxx\, Syy$、つまり相関係数の２乗は「決定係数」と呼ばれ、独自の意味を持つ。計算過程は省くが、この値は $\Sigma(Yi - Y_m)^2 / \Sigma(yi - y_m)^2$（y 予測値 Y のばらつきの、y 値全体のばらつきに対する割合）に一致し、もとのデータの持つばらつきのうち、回帰

直線によってどれくらいの割合が説明できるかの指標となる。たとえば、漁獲と水温の回帰を求めて相関係数 r = 0.8 を得たとすると、$r^2 = 0.64$ となり、単回帰計算の枠組みにおいて、漁獲変動の 64%が水温によって説明されたことになる。ただしこのことは直ちに、漁獲変動要因の 64%が水温であるという直接的な因果関係を示すものではない。相関、回帰と因果の関係についてはのちに述べる。

先に求めた回帰直線の傾き a の有意性は、次のように検定する。もしも回帰による推定 y 値（Yi）の分散 $1/n\,\Sigma(Yi-ym)^2$ が、実際の y 値と推定 y 値の差の分散（誤差分散）、$1/n\,\Sigma(yi-Yi)^2$ よりも十分大きいならば、回帰直線を引く意味があると考える。誤差分散のほうが十分大きいと、回帰直線の傾きは誤差のばらつきの中に埋没してしまい、予測の機能を果たさなくなるだろう。

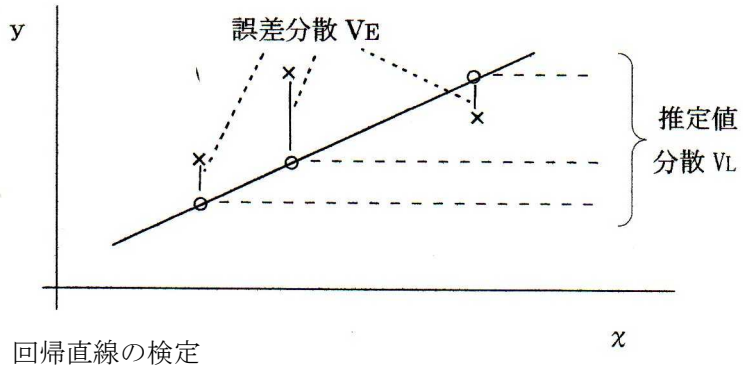

図 2. 回帰直線の検定

そのためには次の値、

$Fcal = V_L / V_R = (Sxy^2/Sxx)/\{(Syy-Sxy^2/Sxx)/(n-2)\}$

を求め、F 分布によって検定する。F 分布の前提は正規分布母集団だから、残差の正規分布が必要条件となる。ここで残差とは、y の Y（＝回帰による予測値）からのズレ（図 2 の V_E）のことである。従って残差の正規分布とは、各点の y 座標の、回帰直線への距離の正規分布であって、y 値そのものの正規分布ではない。またこの他、傾きの信頼限界や 2 つの回帰直線の傾きの差の検定もできるが、それらの計算に当たっても、y 残差の正規分布が前提となる。

２．単相関（ピアソン積率相関）

単相関係数 r は、上記回帰のところでも触れたが、次のように定義される。

$r = \Sigma(\chi i-\chi_m)(yi-y_m) / \{\sqrt{\Sigma(\chi i-\chi_m)^2}\sqrt{\Sigma(yi-y_m)^2}\}$

$= Sxy / (\sqrt{Sxx}\ \sqrt{Syy})$

１行目の式は、以下のように意味づけることができる。

$\chi i-\chi_m$、$yi-y_m$ は、座標平面上にばらつく（$\chi i, yi$）の点を、（χ_m, y_m）を原点とするように平行移動することを意味している（図３i）。この状態で第１象限に入る点では $\chi i-\chi_m > 0$、$yi-y_m > 0$ で、（$\chi i-\chi_m, yi-y_m$）の符号は（+, +）。以下第 4 象限まで、図３ii のような符号関係になる。ここで両者の積 $(\chi i-\chi_m)(yi-y_m)$ の値は第 1、第 2

象限に入る点が多いほど大きく、第２,４象限の点が増えるとマイナスの値に相殺されて小さくなるので、各点が第１,３象限に偏って分布しているかどうかの指標になる。
ただし、このままでは点の数に影響されるからｎで割り、

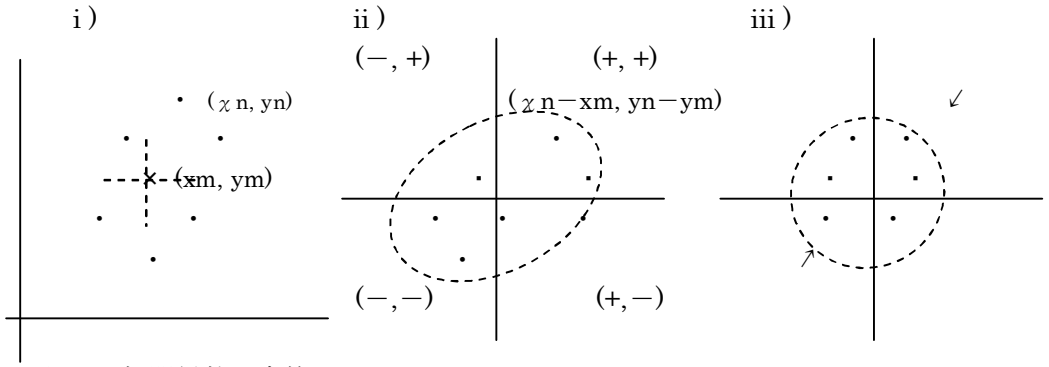

図 3. 相関係数の定義

$1/n\,(\chi i-\chi_m)(yi-y_m)$

これがχとｙの共分散である。さらに、図の矢印方向にばらつきを標準化して、全体の値が−1〜+1 に収まるように、χ、ｙの標準偏差で割ると（図３iii)、

$\{1/n\,(\chi i-\chi_m)(yi-y_m)\}\ /\ \{\sqrt{\Sigma(\chi i-\chi_m)^2/n}\times\Sigma\sqrt{(yi-y_m)^2/n}\}$

$=\Sigma(\chi i-\chi_m)(yi-y_m)/\{\sqrt{\Sigma(\chi i-\chi_m)^2}\,\sqrt{\Sigma(yi-y_m)^2}\}$

で、初めのｒの定義式が得られる。

つまり、相関係数とは標準化した共分散であって、その値が−1〜+1 の範囲に収まる便利さがある反面、もとのデータが持っていたχとｙのばらつきの情報は失われている。このことは、たとえば主成分分析で、計算のベースを共分散行列にするか、相関行列にするかというような問題に反映する。単位系が同じなら、分散の情報を生かせる共分散行列を出発点にするほうが良いといわれるのはこのためである。なお、標準化した形でのχ×ｙの値は、原点中心の同一円周上であれば y＝χの直線上にあるときが最も大きいことが、簡単な計算で確かめられるから、各点が単に第 1、3 象限に入っているだけでなく、同一直線（標準化したときの y＝χ）上にあるときに、相関係数が最も大きいといえる。

この相関係数の有意性は次のように確かめられる。χ、ｙが共に正規分布（２変量正規分布）する前提のもとに、χとｙが無相関な母集団を考え、そこからランダムサンプリングをくりかえすと次のような標準誤差をもったｒの頻度分布が描かれるとされている。

$Sr=\sqrt{\{(1-r^2)/(n-2)\}}$

これもとに次のような F 値、

$F_{cal}=(n-2)/(1-r^2)-r^2$

を計算し、F 分布によって検定できる。このほかｔ値を利用する、求めたｒとサンプル数ｎから直接検定する、ｚ変換という対数を用いた変換によって、ｒを正規分布するｚ値に置き換えて検定する、などの方法もある。上記の標準誤差をもとに、相関係数の信頼限界や 2 つの相関係数の有意差検定も可能である。

ここで登場した 2 変量正規分布とは、χのどの値に対してもｙが、またｙのどの値に対

しえもχが正規分布することを意味し、テキスト類では、χy 平面上に、点頻度がおわんをふせたように盛り上がった図が出ている。相関係数そのものは、ここで示したように定義された指標として機械的に計算できるが、r の有意性や信頼限界を知ろうとすると、母集団の２変量正規分布が満たされていなければならない。生物データのようにこの条件を満たしにくい場合は、Kendall や Spearman のノンパラメトリクス順位相関を使うことも考えられる。

３．回帰と相関

回帰と相関の理論的背景は異るが、計算上の共通性から、しばしば混同して使われている。では両者はどう違うのか、Sokal & Rohlf（1981）に従って解説してみよう。まず、回帰の目的は、説明変数 X によって目的変数 Y の値を予測、説明することであって、通常 X は固定変量、Y はランダム変量である。つまり、X としては特定の離散値しか想定していないが、Y は一応何が出てくるかわからないという形である。場合によっては、X→Y、Y→X の二つの経路の予測、説明がありうるが、その場合、両者の意味づけは異なっている。たとえば、温度に対してある動物の代謝量を打ったグラフで、温度に対する代謝の回帰を求め、ある温度に対する代謝量を予測することができる。逆にある代謝量が与えられたときに、温度はどれくらいだったかを推定することもできるが、固定変量とランダム変量の関係は入れ替わっており、同一グラフ上で引かれる回帰直線も等しくない。なぜ直線が異なるかというと、X→Y の回帰では誤差が Y 軸に平行に回帰直線に引いた線分で表されるのに対し、Y→X の回帰では、誤差が X 軸に平行な線分になるからである。もしも X と Y の値を標準化したあと回帰したり、主成分分析の場合のように、各点から直線に垂線を下ろすのであれば直線は１本に決まり、こうした現象は起こらないが、一般の回帰において X と Y は非対称である。一方、相関を調べる目的は、X と Y が連動して変化しているかどうかの確認であり、X、Y ともにランダム変量、かつ相互に対称の関係にある。実際数式的にも、r の定義においてχと y を入れ替えても同じ式になる。

こうした違いがあるにもかかわらず、回帰分析の枠組みでとられたデータで相関を計算したり、その逆を行うと、しばしば不適切な結果を生む。まず、前者は「無意味」であると言われる。回帰における、「X による Y の予測、説明」という目的は、相関の「両者の連動関係の確認」という目的よりも明確、限定的であり、前者の分析が可能なデータを、後者で分析すれば、話がかえってあいまいになる。ただし、残差母集団の正規分布が明らかでなく、また現象の性質から、一方向的な影響しかあり得ないということなら、ノンパラメトリクスの順位相関などを計算する方がよい場合もあるだろう。一方、相関分析用のデータで回帰を計算することの問題は、「評価の偏り」とされる。相関の目的は対等な２変量の連関評価であり、X と Y の関係を直線で代表させるとすれば、各点からその直線に下ろした垂線の長さを最小にするというイメージになる。しかし回帰直線は、X→Y で引いても Y→X で引いても、共にこの直線と一致せず、直観的に最も妥当な像からずれることになる。

もう一つ、回帰と相関の違いは、前者には残差があるが、後者にはないということである。回帰の場合は X によって Y を説明しようとするから、説明しきれない残りの部分が生じ、それが残差として残る。つまり取り上げた変数の枠内に収まらない、未知の要因を

許容する分析になっている。しかし相関はXとYの枠内ですべてが処理され、他の要因という考え方がない。したがって相関を要因分析に使うときには、常に「要因がこれしかないとすれば」という条件つきである。これは重回帰と偏相関でも基本的に同じといえる。

４．相関と因果

相関を因果分析に使う場合､AがBを支配しているならAとBには相関があるだろう、という発想を前提としている。たとえば、黒潮が近づくと海岸の水温が上がる、という因果関係があれば、黒潮の岸からの距離と水温には相関があるはずである。しかしこれについては、「相関は必ずしも因果関係を反映せず」という、有名なテーゼがある（Shipley 2000）。因果関係があれば普通は相関が出るだろうが、相関のあることが直ちに因果を示すわけではない。「逆必ずしも真ならず」である。これは回帰を因果分析に用いる場合も同じで、回帰係数の有意性はそのまま因果関係の存在を意味しない。

具体的に言うと、AとBに有意な相関が検出された場合、AとBの因果連関については、次のようにいくつかの可能性がある。

１）A → B因果　　２）B → A因果　　３）偶然の相関

４）潜在要因による見かけの相関（C→A因果＋C→B因果で、AとBは無関係、など）

まず１と２だが、AとBの間に相関があるだけでは、AがBを引き起こしたのか、その逆なのかはわからない。ただし、常識的にどちらかに絞れる場合もあるだろう。たとえば、AがBよりも時間的に先行しているとか、黒潮が海岸水温を上げることはあっても海岸水温が黒潮の流路を変えることはありそうもない、など。３は、因果関係はないが、たまたま相関を計算したら有意になってしまったという場合で、これは相関係数の危険率（P< 0.05）によって、評価の中に織り込むことができる。４は最も厄介なケースで、何か別の要因とA、Bが関係しているために、AとBに因果連関はないが、見かけ上相関が現れる。隠れた要因についてはいくつかあったり、経路も複雑になっていることがあり得る。

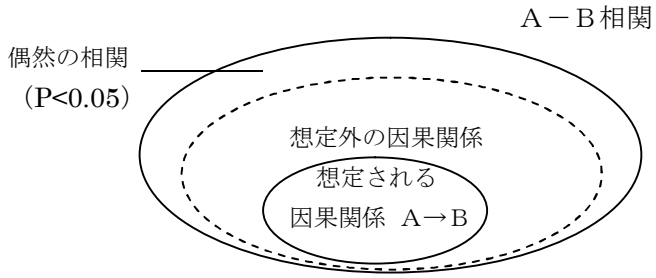

図４．相関と因果の包含関係

相関と因果の関係は、図４のようになる。まず相関のあるケースから、偶然の相関を、危険率によって除く。しかしそれで相関と因果が一致するわけではない。A－B相関の範疇には、A→B因果以外の、B→Aや、C→A＋C→B＝A－Bなどが含まれている。これを排除して初めて、A－B相関はA→B因果に一致することになる。偏相関やパス解析はこ

うした背景要因に視野を広げた分析である。

重回帰（Multiple regression）

１．重回帰の理論

重回帰とはいくつかの説明変数（X1, X2, …Xn）から、線形（＝比例式の和）的に目的変数（Y）の値を推定する方法である。具体的には表２のデータ表を出発点として、

$y\,i = A1\,\chi\,1i + A2\,\chi\,2i + \cdots + An\,\chi\,ni + ei$ …②

（yi, 目的変数; χ i, 説明変数; A, 偏回帰係数; ei, 誤差 or 残差）

という推定式を立てる。変数が一つであれば、この式は

$yi = A1\,\chi\,1i + ei$

となって単回帰を表すから、重回帰は単回帰の一般形といえる。このとき、

$ei = yi-(A1\,\chi\,1i + A2\,\chi\,2i + \cdots + An\,\chi\,ni)$

とし、ei^2 を最小にするように A1, A2… の係数列を決定することができれば重回帰式が求まる。これは、多次元空間に最小２乗法で回帰直線を通す作業といえる。そのためには単回帰の場合と同様、

$ei^2 = \{ yi-(A1\,\chi\,1i + A2\,\chi\,2i + \cdots + An\,\chi\,ni)\}^2$

が、A1, A2… のそれぞれの２次式になることをもとに、A1, A2… それぞれに対する偏微分がゼロになるように偏回帰系数列を決定する。計算過程は行列方程式を利用した大がかりなものになるので、ここでは示さないが、結果としての偏回帰係数の一般形は次のようになる。

$Ai = S^{i1}\,S_{1y} + S^{i2}\,S_{2y} + \cdots + S^{in}\,S_{ny}$　ここで、

Ai：重回帰式における i 項偏回帰係数

S_{ny}：説明変数 χ nj と目的変数 y の偏差積和　$\Sigma(\chi\,nj-\chi\,nm)(y\,j-ym)$

S^{in}：$\chi\,1$～$\chi\,n$ の偏差積和 n 行 n 列正方行列式を求めたときの、i 行 n 列対応要素

特に２変数の重回帰式

$y = A1\,\chi\,1 + A2\,\chi\,2$　では、

$A1 = (S_{1y}\,S_{22}-S_{2y}\,S_{12}) / (S_{11}\,S_{22}-S_{12}^2)$ ･･･③

$A2 = (S_{2y}\,S_{11}-S_{1y}\,S_{12}) / (S_{11}\,S_{22}-S_{12}^2)$　となる。

パソコンプログラムでは、表２のデータを打ち込んで重回帰を選択すれば、必要な数値が自動的に output される。こうして重回帰式、

$Yi = A1\,\chi\,1i + A2\,\chi\,2i + \cdots + An\,\chi\,ni$

が決まると、説明変数列 χ 1i, χ 2i, χ 3i … にデータを代入して Yi（予測値）を求めることができる。たとえば、水温と塩分と流速がこれこれの時、漁獲量はこれくらいだろう、といったぐあいである。

このとき、重回帰式からの予測によって得られた Yi と誤差を含んだ実際の値 yi は、通常いくらか異なっている。両者の相関係数

$R = \Sigma(Yi-Y_m)(yi-y_m) / \{\sqrt{\Sigma(Yi-Y_m)^2}\,\sqrt{\Sigma(yi-y_m)^2}\}$

を重相関係数といい、回帰式のあてはまりの良さを示す指標となる。複数の説明変数と単一の目的変数との関係という意味で、「重」相関というのであろう。相関係数の性質とし

てRは−1〜1の値をとるので、R^2は0〜1の値となり、これを「決定係数」と呼ぶ。単回帰では相関係数の２乗を決定係数と呼んだが、それに相当する概念であり、この場合もまた、重回帰式によって、目的変数のばらつきのうちどれくらいの割合を説明できるかの指標となる。

ここで注意すべきこととして、この重相関係数は各説明変数と目的変数の間の関係の強さの他、単純に、回帰に用いた説明変数の数にも影響され、説明変数の数が、ケース数−1になった時点で必ず1になるという奇妙な性質がある。目的変数と明らかに無関係な指標でも、どんどん加えていけば100%の予想が可能（？）であるかのようなこの結果は、重回帰計算の理論的帰結である。つまり重回帰とは、表中の説明変数を重みづけしてたし合わせ、目的変数に近づけて行こうという単なるアルゴリズムであるから、各説明変数が背後にどのような意味を抱えているかに無関係に、最も当てはまりの良い係数列を打ち出してくる。それが「ケース−１」で飽和に達するということである。そこでこれへの対策として「自由度調整済重相関係数」というものが考えられており、目的変数との関連の小さい説明変数をいくら加えても、相関係数が高くならないように調整される。

以上のような手順で重回帰式が求まると、そこに現れる偏回帰係数の列A1, A2… は、対応する説明変数の、目的変数に対する結びつきの強さを示すように見える。Aが大きいほどAχも大きく、Yの中に占める割合が増すからである。しかしこの値は、説明変数χの単位系や、始点からの距離にも影響される。たとえば、重さをmgとgで表すのとでは、同じ現象に対して回帰直線の傾きが1000倍異なる。しかし、だから効果も1000倍ということではない。薬品の投与量をmgで測った時とgで測ったときで、生物の反応が異なるなどということはないからである。同様に、温度を絶対温度で測るのとセ氏温度で測るのとでは、273°の差が出る。こうしたことを補正するために、各説明変数値の平均値からの差をとり、さらに標準偏差で割った上で重回帰式を求める。このときの偏回帰係数を標準偏回帰係数といい、この形において初めて、目的変数に対する各説明要因の相対的寄与を評価できることになる。なお、標準偏回帰係数は、説明変数間の相関がゼロの場合、各説明変数の、目的変数に対する単相関係数に一致する（奥野ほか1971）。従って、この値の2乗はおなじみの決定係数であり、目的変数のばらつきのうち、当該説明変数が担う割合を示す指標となる。

重回帰式の、全体としての当てはまりのよさは、重相関係数によるほか、単回帰の場合同様、「回帰による分散／誤差分散」を指標とするF検定によって調べることもできる。この他、標準誤差に基づく偏回帰係数の有意性や信頼限界などの評価も可能だが、単回帰同様、誤差分散の正規分布が前提になる。

２．説明変数の選択

重回帰式に、どのような説明変数をいくつ取り入れるかということが、しばしば問題になる。それを決めるための方法はいくつかあるが、ここでは一般のパソコンソフトに最もよく取り入れられているステップワイズ法を紹介する。

利用可能な説明変数のうち、目的変数との単相関係数の有意性が最も高いものを選び、その説明変数で単回帰式を作る。次に残りの変数のうち、回帰式に取り入れて重回帰式を構成したとき、その偏回帰係数の有意性が最も高いものを加えて、２説明変数の重回帰式

を作る。同様にして、残された未検討変数から、重回帰式に取り入れたときに最も偏回帰係数の有意性が高いものを順番に加え、説明変数を増やしてゆく。ただしこの場合、新たな変数が式に加わると、既に存在し、かつては有意だった説明変数が非有意になることがある。この場合はその変数を重回帰式から除き、未検討変数群に戻して、その中からまた最も有意性の高い変数を探すという操作を繰り返す。このようにして、加えたり減らしたりをくり返しながら、最終的に、重回帰式の中にあるすべての説明変数の偏回帰係数が有意であり、かつ残りの変数のからどれを加えても有意にならないという状態に落ち着くと、最終的な重回帰式が決定する。以上は「増加型ステップワイズ」と呼ばれるやり方だが、利用可能なすべての変数から出発して減らしたり増やしたりしながら重回帰式を導く、「減少型のステップワイズ」もある。このほか、前進後退をくり返さず、一方的に増やしてゆくのが「前進型」、はじめにすべての変数を取り込み、一方的に減らしてゆくのが「後退型」で、これらのうちいずれをとるかで、最終的な重回帰式が異なることもあると言われている。

３．重回帰の条件

重回帰の目的が、与えられたいくつかの説明変数から、目的変数を推定する際の、最も当てはまりのよい式を得るだけならば、以下に述べる諸々の条件は必要ない（フラーリー・リードウィル 1990）。既に述べたような定義によって与えられた式であることを承知した上、機械的に目的変数の値を計算することができる。単に、直観的に判断するよりはましな推定値を得たいというほどのことならそれで足りるし、そのような用法が適切な分野もあると思われる。しかし偏回帰係数の有意性、信頼限界や、係数どうしの差の有意性などを評価しようとすると、母集団にかかわる以下のような条件が必要になる（塩谷 1990）。

１）説明変数（y）と目的変数（$\chi 1, \chi 2, \chi 3$ …）の関係が線形的である。

２）各説明変数の組に対する誤差は、互いに独立。

３）誤差は、平均ゼロで、ある分散を持った正規分布に従う。

４）目的変数どうしは、相互に相関しない。

１）の線形性は、すでに述べたように、目的変数が各説明変数の比例式の和で表されることで、式②（前々頁）の形を意味している。しかしこの前提は、生物データに対してはしばしば非現実的である。たとえば一般に、生物の代謝は温度に対して指数関数的に増加するし、時間と個体数変化のロジスティック式では、初めは指数関数的、密度が飽和に達すると頭打ちの形となる。薬剤 $\chi 1$ と $\chi 2$ の相互作用なども十分考えられる。あるいは反応に特定の境界値があり、その値を越えると急激な変化が起こるというようなのも線形ではない。指数項（αp^{χ}）や累乗項（$\alpha \chi^{n}$）を組み込んだ非線形回帰や、相互作用項（$\alpha \chi 1 \times \chi 2$）を含む式も考えられているが、なおカバーできない関係もあり、どういうパターンなのか実際のところはわからないなどとなればお手上げであろう。

２）と３）は、パラメトリクス統計の定番の要件である。偏回帰係数の検定は分散比に基づく F 検定で行われるから、ランダムサンプリングと誤差母集団の正規分布が前提となるのである。

４）はいわゆる「多重共線性」（multi-colineality、略称マルチコ）の問題である。採用した説明変数間に強い相関があると、結果が不安定になったり、算出した偏回帰係数間に

相互の影響が混入し、それぞれの説明変数が独立の効果を代表するとはいえなくなる。なぜそうなるのか、最も単純な２変数の重回帰式、$y = A1\chi 1 + A2\chi 2$　で説明すると次のようになる。先の重回帰理論の③式のように、偏回帰係数 A1, A2 の分母は、$S_{11} \cdot S_{12} - S_{12}^2$ である。一方、変数 $\chi 1, \chi 2$ との相関係数の 2 乗は $r_{12}^2 = S_{12}^2 / S_{11} \cdot S_{22}$ だから、これが 1 のとき、$S_{12}^2 = S_{11} \cdot S_{12}$ から、$S_{11} \cdot S_{12} - S_{12}^2 = 0$ となる。分母が 0 になれば偏回帰係数は計算できず、また相関係数が 1 に近いところでは、データのわずかな差によって偏回帰係数が大きく変動する。また、S_{12} は２変量の共分散であり、相関が高いとこの値が大きくなる。S_{12} は分母と分子の両方にあるのでその影響は単純ではなく、場合分けを伴ったかなり複雑な話になる。ここでは具体例に即して結果だけを定性的に示す。今、都市の人口、工場数を説明変数、その都市の大気汚染度を目的変数として重回帰分析を行うとすると、人口、工場数とも汚染源になりうるから、それらの標準偏回帰係数はプラスになると予想される。しかし工場の寄与が人口より大きく、かつ人口の多い都市はベッドタウン的性格を持っていて工場数が少ない、というような負の相関があると、人口の項の係数はマイナスになることがある。このときこの結果から、人口が多いほど汚染が少なくなると結論すればナンセンス、ないし不正確な表現となろう。

多重共線性への対策をめぐっては、２つの考え方がある。一つは、説明変数間の相関は、各説明変数の要因としての関与の理解を妨げるから、下調べの上、相関する要因についてはあらかじめどちらかを除いて重回帰式を作るというものである。分析を重回帰の枠内に止めようとするとこういう方向性になるのだろうが、しかしこれをやると、当然のことながら除かれた要因は評価できなくなる。生物データでは、要因相互の相関は普通である。カニの交尾成功率に対する体サイズと鉗脚サイズの寄与を評価しようとすれば、後２者の相関は明らかだし、魚の漁獲に対する水温と塩分の関与を調べるにしても、それらの背後には黒潮など海況の影響がある。もう一つは、一見不合理な偏回帰係数が得られたら、なぜそうなっているのかを考え、背後の相関関係をも含めて現象を再評価することである。これは次に述べる偏相関の考え方であり、さらに発展するとパス解析になる。

偏相関（Partial correlation）

１．偏相関の理論

偏相関は、複数の変数が相互に連関しているとき、特定の二つの変数の関係を、他の変数の影響を除去して評価したい場合に用いられる。線形式を理論的基礎にしているので重回帰との共通点は多いが、偏相関といえども相関の一種であり、重回帰は回帰だから、先に述べた相関と回帰の違いはこの場合にもあてはまる。つまり偏相関は対等な二つの変数の連動関係を見て双方向的、重回帰は複数説明要因による特定の目的変数の予測、説明を主眼として一方向的、ということである。

偏相関係数の導出過程は次のようになる。おなじみの線形式

$$y = A1\chi 1 + A2\chi 2 + \cdots + An\chi n + \varepsilon$$

において、y と $\chi 1$ の関係を $\chi 2 \cdots \chi n$ の影響を除いて考えてみる。なお、わかりやすさを考えて重回帰と同じ式を使っているが、相関の性質上 χ と y は対等で、意味づけに特別

な差があるわけではない。上式においてｙと、χ１を除いた残りχ２…χｎの回帰を求めると

$y = A2\chi 2 + \cdots + An\chi n + \varepsilon'$

平均からの偏差をとって

$y - y_m = B2(\chi 2 - \chi_m) + B3(\chi 3 - \chi_m) + \cdots + Bn(\chi n - \chi_m) + \varepsilon_y$

残差は　$\varepsilon_y = (y - y_m) - \{B2(\chi 2 - \chi_m) + B3(\chi 3 - \chi_m) + \cdots + Bn(\chi n - \chi_m)\}$

一方、χ１をχ２,χ３ … で説明する回帰式も

$\chi 1 = C2\chi 2 + C3\chi 3 + \cdots + Cn\chi n + \varepsilon$ ” をもとに

$\chi 1 - \chi_m = C2(\chi 2 - \chi_m) + C3(\chi 3 - \chi_m) + \cdots + Cn(\chi n - \chi_m) + \varepsilon_{x1}$

残差は　$\varepsilon_{x1} = (\chi 1 - \chi_m) - \{C2(\chi 2 - \chi_m) + C3(\chi 3 - \chi_m) + \cdots + Cn(\chi n - \chi_m)\}$

そしてこのとき、ε_yとε_{x1}の相関が偏相関係数となる。つまりこれは、χ１を除くχ２, χ３ …でｙを回帰したときの残差ε_yと、χ１をχ２, χ３ …で回帰したときの残差ε_{x1}の相関であって、理論的に、χ２, χ３ … の影響を除いたｙとχ１の相関を示していることになる。χ２,χ３ … の影響は、それらでｙとχ１を回帰した段階で除かれたと考えるのである。このイメージは、図４のようになる。

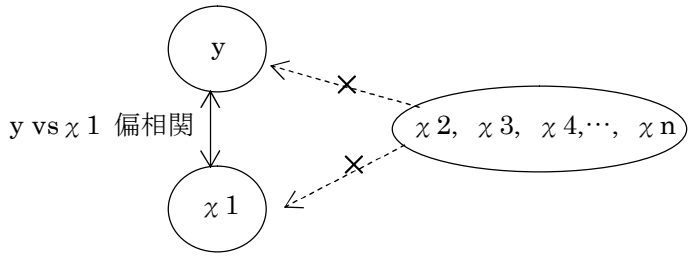

図５．変数ｙとχ１の偏相関

ε_yとε_{x1}の相関係数の実際の計算は、行列、逆行列を利用した大がかりなものになる。最終的な結果だけ示すと、最も単純な、変数がｙ,χ１,χ２の３つの場合、χ２の影響を除いたｙとχ１の偏相関は、

$r_{x1y.x2} = (r_{x1y} - r_{x1x2} \times r_{x2y}) / \sqrt{\{(1 - r_{x2y}^2)(1 - r_{x1x2}^2)\}}$　…④

つまり、ｙ、χ１、χ２の、相互の単相関係数から計算できる。

ここで、χ２がｙ, χ１になんら影響を及ぼしていない、つまりχ２とｙ, χ１の相関が０のとき、$r_{x1y.x2} = r_{x1y}$となり、偏相関は単相関に一致することを確認しておく。変数が多い場合、ｙ,χ１,χ２,χ３ …においてχ２,χ３, …χｎの影響を除いた、一般形としてのｙとχ１の偏相関は、

$r_{x1y.x2} = -r^{1y} / \sqrt{(r^{11} r^{yy})}$

ここにr^{1y}は、ｙ, χ１,χ２,χ３, …,χｎの相互の単相関を総当りで求めたときの（ｎ＋１）行（ｎ＋１）列相関行列式において、その逆行列の（χ１, ｙ）対応要素を示す。この式を、３変数の場合同様、ｙ,χ１,χ２,χ３,…,χｎ間の単相関係数を使って表すこともできるが、極めて煩雑になる。

２．偏相関の条件

偏相関においては、対象とする２変数が上に述べたように、他の変数によって線形的に表現されることが前提になっている。また、偏相関係数の有意性や信頼限界を求める場合、その検定は単相関の場合のように F 検定（または t 検定）で行われるから、各変数残差の、多変量正規分布が満たされねばならない。

一方、重回帰で問題になった多重共線性は気にしなくてよい。というよりも偏相関では、変数間に相関があることはむしろ前提である。したがって、生物データのように、要因どうしの相互連関が頻繁に起こる分野では、因果関係分析の有力な手段となりうる。たとえば、先に出てきたカニの体サイズと鉗脚サイズの、交尾成功率に与える影響、といったテーマでは、体サイズと鉗脚サイズの相関は明らかだから重回帰による要因推定は不適切だが、偏相関は使える。この場合、偏相関は理論上、双方向的な連動関係を示し、説明変数→目的変数の一方的な関係を示すものではない。しかしこの例では、交尾成功率→体サイズの逆方向の因果は少なくとも短期的にはありえないから、回帰と同様の説明的効果を持つと考えてよい。

ただし問題なのは、偏相関は選択した要因の枠内に限定された分析だということである。そこに新たな変数を加えると前の枠組みがこわれ、その枠内で有意だった変数が有意でなくなったり、その逆も起こりうる。単純な例を挙げると、$\chi 1$ と y の単相関 $r_{x1,y}$ が正の値（たとえば 0.5）であっても、第３要因を加えた偏相関の定義式（上記④式）において、$\chi 1$ と $\chi 2$、$\chi 2$ と y の単相関値が大きい（たとえば 0.7 と 0.8）ならば、分子 $r_{x_1 y} - r_{x_1 x2} \times r_{x2y} = 0.5 - 0.7 \times 0.8 < 0$ となり、分母は必ず正であるから、単相関と偏相関の符号が逆転する。従って実験室内のように厳密に条件をコントロールできるとか、野外であっても、事実上これしか要因は考えられない、という場合以外は、偏相関による検討は暫定的なものに止まらざるをえない。こうしたことは、上記偏相関の導出過程からも明らかだろう。y と $\chi 1$ の偏相関の計算では、y については $\chi 2, \chi 3 \cdots \chi n$ で説明できない変動はすべて $\chi 1$ に由来し、$\chi 1$ についても $\chi 2 \cdots \chi n$ で説明しきれない変動は y に由来することを前提にして、相互に残差の相関を求める。はじめから $y, \chi 1, \chi 2 \cdots \chi n$ 以外の要因は想定されていない。

残差の有無は、重回帰と偏相関の重要な相違点である。重回帰の場合は残差があり、採用した説明変数で説明しきれない分はこの中に含めて考えるが、偏相関に残差はない。そもそも残差とか未知要因というものは、何かで何かを説明しようとした残りの部分ということで、方向性を前提としている。しかし偏相関は相関の一種で双方向的だから、概念的にも残差というものは考えられないのである。

３．各係数の関係

ここで、これまで出てきた相関、回帰にかかわる諸係数の関係をまとめておこう。

単相関と単回帰では、すべての点が一直線上に乗るとき、単相関係数は単回帰係数に一致する。

単相関と偏相関では、後者において当該２変数以外の変数の、これら２変数との相関が０であるとき、単相関係数と偏相関係数は一致する。

単回帰と重回帰では、重回帰における偏回帰係数が、説明変数間無相関の条件下で、目

的変数と各説明変数の単回帰係数に一致する。

単相関と重回帰では、重回帰における標準偏回帰係数が、説明変数間無相関（ないし説明変数が 1 つのみ）の条件下で、目的変数と各説明変数の単相関係数に一致する。

偏相関と重回帰では、偏相関係数と偏回帰係数の間に、次のような関係がある（奥野ほか 1971）。

（y vs χ1 の偏相関係数)^2 =（y→ χ1 の偏回帰係数)^2 ×（ χ2←y の偏回帰係数)^2

つまり、2 変数の偏相関係数は、それらを相互に重回帰したときの偏回帰係数の幾何平均である。偏相関では、重回帰の方向性が、相互のかけ算によって相殺されているわけで、「回帰→方向、相関→双方向」の性格が、数式的にもよく表現されている。

パス解析（Path analysis）

パス解析は現在発展中の分野で、分析や検定にもいろいろなパターンがある。しかし基本的には重回帰の積み重ねと考えてよいと思われる。ここではこれまでの記述を生かし、重回帰を階層的に用いる古典的方法を紹介する。

パス解析においては、まず、因果連関の推定図（パス図）を描く。図 5 のように、目的変数 Y に対して X1 と X2 が、さらに X2 に対して X3 が関与していると考えたとする。図中の U, V は未知要因（unknown factor）を示す。

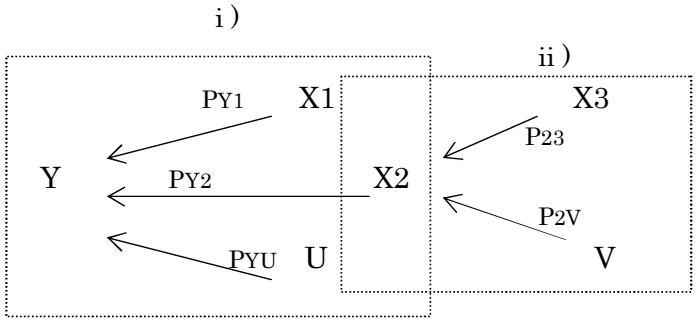

図 6．パス図の例

図の矢印に付した p はパス係数と呼ばれ、矢印の経路の相対的重要性を示す指標になる。これを決定することがパス解析の一つの目標である。まず（ii）の点線枠内において X2 に対する X3 の回帰（この場合は単回帰）を考え、標準化した回帰係数（＝相関係数）を p_{32} とする。これは X2→X3（＝X3 を X2 で説明する際）のパス係数を意味し、慣例的に目的変数を先、説明変数を後に置く。また、相関係数の 2 乗は決定係数だから、これによって残りの分散が表現されていると考えて $p_{2V} = \sqrt{(1-p_{32}^2)}$ と決める。（i ）の枠内は重回帰となる。同様に Y に対する X1、X2 の重回帰を考え、標準偏回帰係数を計算してそれらを p_{Y1}、p_{Y2} とする。つまり、パス図において各要因の寄与を示すパス係数は、基本的に重回帰の標準偏回帰係数である。全分散のうち、X1、X2 でカバーしきれない部分（重回帰式の残差部に相当）が p_{1U} で、$\sqrt{(1-p_{Y1}^2-p_{Y2}^2)}$ によって算出する。

パス図（モデル）の妥当性を評価するには、重回帰で得られたパス係数を、回帰と相関の理論に基づき、パス図に付与された様々な性質（テキスト類に出ている）を使って検算する。たとえば X1→ X2→ Y の経路を考え、パス係数 p_{21}、p_{Y2} が得られたとして、その図が正しければ、X1 と Y の相関 $r_{1Y} = p_{21} \times p_{Y2}$ とならなければならない。あるいは変数 Y に対し X1 と X2 の関与のみであれば、パス係数どうしの関係は、$r_{Y1} = p_{Y1} + p_{Y2} \times r_{12}$ となる。いずれにせよこうした検算によって、両辺の値に無視できない差が生じた場合は、パス図が不適切と考え、説明変数を加えるなど、修正しなければならない。

パス解析は重回帰を使うが、多重共線性は問題にしない。説明変数間のそうした連関を前提として、背後の要因や、それらを含めた全体像を評価しようとしているからである。パス係数（標準偏回帰係数）が正になるはずだが実際に負になっていれば、それは背後に図のような連関があるからだろう、と考える。その意味でパス係数は、常に個々の説明変数それ自体の、目的変数への純粋な関与の度合いを示すというわけではない。一方変数間の線形性や残差の正規分布は、パス係数の有意性を問題にする限り依然として避けられない条件である。

パス解析は現在、教育学、心理学など社会科学分野で多用されているが、これらの分野では共分散構造分析（または構造方程式モデリング）と呼ばれる手法が主流になりつつあるらしい（豊田ほか 1992）。これは、観測データから成る「観測変数」以外に、それ自体は具体的データのない構成概念としての「潜在変数」（社会的地位、性格、知能、など）を導入し、それらを組み入れた重回帰式によってパス図を描く。潜在変数の抽出は観測データからの因子分析による。検定も、上で示したような個々の経路についての手作業的なやり方から進んで、全体的なパス図の適合性を、カイ 2 乗検定や情報理論による適合度評価（AIC）によって行う。パス図を描き、変更しながら繰り返し検定して、最も観測データとの適合性のよい経路を構成するためのプログラムも開発されている。共分散構造分析においては、観測データ自体、背後に何らかの潜在変数（真の値）を持つと考える。たとえば「温度」をアルコール温度計で測れば、かならず真の値からの何らかの誤差を含むという意味で、「温度」なるものは潜在変数である。こうした考え方は正確といえば正確だが、これらによってパス図はますます複雑化し、結論があいまいになる傾向は否めない。

分析の実例

ここでは具体的なデータをもとに、重回帰、偏相関、パス解析の計算を試行してみる。例として、私のかかわった、和歌山県白浜番所崎の海岸貝類相の年変化と環境要因の問題を取り上げる。当地では、1985 年から貝類調査が行われており、年々群集中の南方性の要素が増加しつつある（Ohgaki et al. 1999）。南方性種は、冬季の低水温の影響を受けやすいと考えられ、実際調査地周辺では、寒波など低水温時に熱帯性種が大量死することが知られている。そこで、目的変数として貝類相中の、南方性種の優占度（房総以南に分布が限られる種の出現頻度の、全種出現頻度累積値に対する割合、以下 S%と略）をとり、説明変数として 2 月の平均海岸水温を考える。他に海岸貝類に影響する要因として気温、海岸水温に影響する要因として沖合水温、さらに沖合水温を支配する可能性のある要因として、黒潮の紀伊半島からの離岸距離を取り入れる。このうち S%は当年データだが、環

境要因は累積的に影響すると考え、貝類調査の行われた年とそれに先立つ 2 年、計 3 年分を平均して用いた。各値は 1985～2000 年の 16 年度分、データ表の概略は表 3 のようになる。なお、特に目的変数は生物データかつ百分率なので、誤差分散の正規性など、重回帰分析の前提を満たさない可能性もあるが、方法の概略を示すのが目的なので、ここでは詳しく検討しない。また、リアリティを重視して実際のデータを使うが、分析は暫定的なものであり、今後の補充や修正もありうるので、導かれる結論自体に現時点では責任が持てないことをお断りしておく。

目的変数		説明変数（環境指標、2 月）			
年	S %	海岸水温	気温	沖合水温	黒潮距離
1985	34.3	12.8	6.2	15.9	60km
1986	32.0	13.1	5.4	15.9	56
:	:	:	:	:	:
:	:	:	:	:	:
2003	46.0	15.1	7.5	17.0	52

表 3. 貝類相変動分析のデータ表。

まず、分析の中心となる海岸水温と S%の関係を見る。水温に対する S%の回帰式は、

Y (S%) = 4.88 X (海岸水温)－31.5

標準(偏)回帰係数（この場合は説明変数 1 つにより単相関係数に一致）は 0.826 で有意 (P< 0.01)。決定係数は $R^2 = (0.826)^2 = 0.683$ で、S%の全分散の 68%を海岸水温で説明できる。つまり、過去 3 年の冬季水温が高いと貝類相中の南方性種の割合が上がるという、かなり強い連関が認められる。残差は$\sqrt{(1-0.68)} = 0.66$ となる。なお残差は一般的に複数要因の複合体だから、有意水準の概念はない。

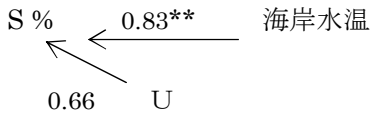

図 7．貝類相 S%と海岸水温の単回帰分析。数字は回帰係数、U は残差（未知要因）。

S%と海岸水温の関係はわかったが、潮間帯は大気にもさらされている。気温の影響はどうなのか。そこで次に、環境要因に気温を加えて、2 説明変数による重回帰分析を行う。その結果は、

Y (S%) = 5.11 X1 (海岸水温)－0.34 X2 (気温)－32.37

各要因の重みを評価するため標準偏回帰係数を求めると、X1 に対して 0.865 (P< 0.01)、X2 では－0.06 (ns)で、海岸水温のみ有意。気温の寄与は小さくほとんど影響していないと、一応推測される。残差（未知要因）の寄与率は 1 から各要因の決定係数を差し引いて$\sqrt{(1-0.865^2-0.06^2)} = 0.867$ と計算される。以上により、各要因の S%との関係は下図

のようになる。

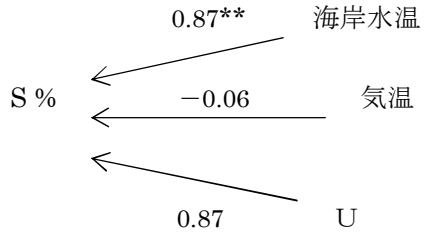

図８．S%と海岸水温、気温の関係。未知要因以外の矢印の数字は標準偏回帰係数。

ここで注意しなければならないことは、気温は海岸貝類だけでなく、水温にも影響する可能性があることである。実際、気温と海岸水温の単相関を取ると r = 0.630（P<0.01）となり、有意な相関がある。つまり上記の重回帰式では多重共線性が発生していると考えられるから、求めた標準偏回帰係数を海岸水温、気温の S%への直接の寄与度に読み替えることはできない。そこで、パス解析とは直接関係ないが、少し寄道をして偏相関を調べてみよう。海岸水温と気温の相関を考慮した上、海岸水温と S%、気温と S%の偏相関を計算する。結果は図９のようになり、やはり海岸水温だけが S%に対して有意であることがわかった。

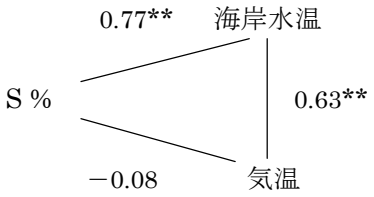

図９．３要因間の偏相関分析。数字は、水温－気温間のみ単相関係数、他は偏相関係数。相関のため、線に矢印（方向性）がないことに注意。

ここで図８に戻ると、ここには２つの問題が見て取れる。一つは今述べたように、気温から海岸水温への経路が考えられていないこと。もう一つは、S%に対し、海岸水温に匹敵する支配要因が、まだ未知要因として隠れているらしいことである。まず一つ目について、海岸水温を支配する要因は何なのか、気温以外にも視野を広げて考えてみよう。

海岸水温に対し、気温以外に影響を及ぼす要因として、最も考えやすいのは沖合水温である。そこで今度は海岸水温を目的変数とし、沖合水温と気温を説明変数とする重回帰分析を行う。結果は以下、

Y (海岸水温) = 0.31 X1 (気温) + 0.99 X2 (沖合水温)－4.39

標準偏回帰係数は、X1, 0.328 (P< 0.01)、X2, 0.717 (P< 0.01)で、共に有意。残差の寄与は$\sqrt{(1-0.328^2-0.717^2)} = 0.63$ と示される（図８）。

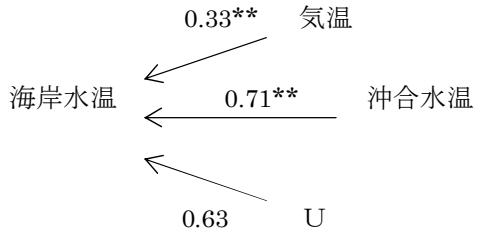

図 10. 海岸水温についての重回帰分析。数字は図８説明参照。

海岸水温に対して気温が有意となり、先の偏相関とは異なる結果が得られた。一方、沖合水温の影響はやはり大きいらしい。さらに、沖合水温に影響する要因を考えてみる。調査点が位置する紀伊半島の西南岸では、一定範囲内であれば黒潮が岸に近づくと沿岸の水温が上がり、遠ざかると下がるという、比較的単純な関係が知られている。とすれば、黒潮の離岸距離と沖合水温は逆相関することが予想される。両者の回帰式は、

Y (沖合水温) = －0.03X (黒潮距離) + 18.11

回帰係数は一見小さいが、これは単位の問題が絡むからセオリー通り標準化し、－0.658（P< 0.01）を得る。予想通り、黒潮距離が小さい（黒潮が岸に近づく）ほど沖合水温が上がるという有意な関係がある。ちなみに残差の寄与は、$\sqrt{(1-0.658^2)} = 0.66$（図９）。沖合水温に対し、黒潮に匹敵する未知要因が残されているようだが、この問題はこれ以上追及しない。

以上の部分的回帰、重回帰分析を総合して、図 10 の全体図（パス図）を得る。ここでは説明の都合上、部分から全体へ分析を広げていったが、実際には初めにパス図を描き、各部分の分析を行うことの方が多いだろう。

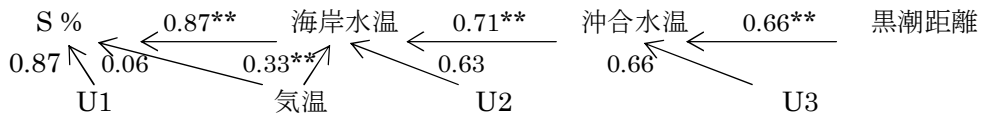

図 11. 貝類相 S%と環境要因についてのパス解析の結果。U1～U3, 未知要因。

この図の妥当性はどうだろうか。パス解析の理論によれば、パス図中の離れたところにある 2 つの項目の単相関は、両者を結ぶ経路上の各パス係数のかけ算で表される。そこで黒潮距離と S%を見ると、このパス図で、黒潮は→沖合水温→海岸水温→S%、という経路のみを考えているので、－0.66×0.71×0.83 =－0.39 と計算される。一方、S%と黒潮距離の単相関は r =－0.72 で、意外に高い。これは、このパス図のどこかに問題があることを思わせる。先に、S%に影響する要因として、海岸水温以外にまだ重要なものが隠れているのではないかと予想した。南日本の太平洋岸では、黒潮の幼生供給によって熱帯性種の個体群が維持されている例が知られている。もしも調査点においても、黒潮が南方性種の幼生を供給することによって S%を上げるという直接的効果を及ぼしているとすると、図 11 において、黒潮から S%への経路を設定する必要がある。そこで、その経路を取り入

れて図を作り直してみる。方法はこれまで述べてきたものと基本的に同じなのでくり返さないが、結果は図 12 のようになる。ここまでくるとかなりパス図らしくなってくる。

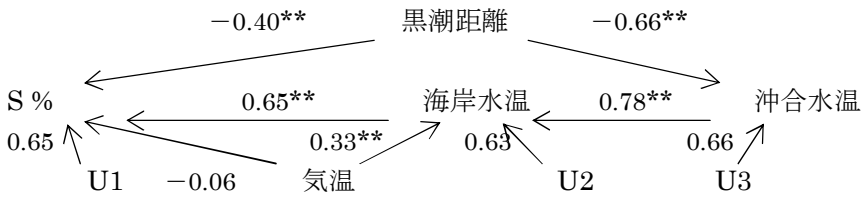

図 12. 貝類相 S%と環境要因。修正されたパス図。

S%に対する黒潮の標準偏回帰係数は−0.40（P< 0.05）で有意となる。ここで再び、黒潮と S%の関係を検討すると、黒潮距離→沖合水温→海岸水温→S%の間接的経路と、黒潮距離→S%の直接的経路のパス係数（の積）をたし合わせることにより、−0.66×0.78 × 0.65−0.40 =−0.74 を得る。これは先に示した、黒潮と S%の単相関係数−0.72 とかなり近く、黒潮と S%を直接結びつけることにより、モデルの適合性が向上したことを示している。

まとめ

ある現象に対して何らかの要因を原因として予想したとき、前者を目的変数、後者を説明変数として単回帰を調べる、もしくは両者の単相関をとるというのが、もっとも単純なやり方である。このとき、相関は双方向の連動関係なのに対し、回帰は一方向の予測であり、未知要因を残差に含めて評価できるから、基本的には回帰のほうが得られる情報は多い。この場合前提条件として、回帰では変数間の線形性や残差の正規分布、相関では 2 変量正規分布がある。これは、回帰係数や相関係数を単なる指標値や予測のための道具として使うのであれば必要ないが、係数の有意性や信頼限界を求めるためには要件となる。そこでこれらを満たしにくい生物データの場合、方向性の仮定の下に、Spearman や Kendall のノンパラメトリクス順位相関を用いることがしばしば行われている。

説明変数が複数あるときには、単回帰は重回帰、単相関は偏相関に発展する。重回帰では、標準偏回帰係数の比較によって、各説明変数の、目的変数への寄与度を評価でき、また未知要因も含めた総合的分析が可能である。しかし先に示した回帰の前提条件に加え、多重線形性の制約により、生物データへの適用はしばしば困難となる。偏相関は要因間相互作用を前提とするので多重共線性の問題は発生しないが、多変量正規分布の制約はある。また残差の概念がなく、採用された変数系の枠内の議論に止まらざるを得ない。パス解析は残差を含み、かつ多重共線性を許容して、その背後の要因にも迫ろうとするもので、いわば重回帰と偏相関の弱点をカバーする両者の発展形態といえる。しかしパス解析も、重回帰の標準偏回帰係数をパス係数として取り入れるので、その有意性評価に当っては、線形性や誤差分散の正規性の制限を免れない。

方法が発展するのは、前段階の手法における欠点をカバーするためであるが、その一方

で複雑化に伴うデメリットも発生する。単回帰、単相関のように分析が単純、単階層に止まる場合、不確定要素も小範囲に収まる可能性があるが、パス解析のように複雑、多層的になってくると、不確定要素が積み重なって、全体としてどの程度のバイアスをもたらしているか、判断しにくくなる。最近では、パス図の適合性を全体的に評価するのが普通だが、適合度が高ければ真実に近いとも限らず、バイアスの積み重なりや打消し合いによって、たまたまそうなっているとか、決定的な未知要因が隠れているのではないかという疑念をぬぐうことができない。違った経路のパス図でも、同じような適合度が得られることがあるとされている（豊田ほか 1992)。直観的に言うと、要因を増やし経路を複雑化させることはある側面から見れば精密化だが、別の側面では曖昧さが増し、両者の効果が相殺して一方的に信頼度が上がっているという感じがしない。たとえば実例検討の項で示唆された、黒潮による海岸貝類相への直接の幼生供給の件にしても、そのほうがパス図の適合がよくなるから存在するにちがいないと言えるのかどうか。この結果は、たとえば何かの熱帯性種について、分布北限での個体群動態や生殖腺成熟、近隣個体群との遺伝組成の比較などにより、この種の個体群が黒潮による幼生供給に依存すると判定されたという場合ほどの信頼度があるとは思えない。後者は一次資料として引用可能だが、前者は無理であろう。そしてこの曖昧さということが、生態学の論文においてパス解析をあまり見かけないことの一つの理由のように思われる。

重回帰式やパス図の中にどのような変数を取り入れるか（仮説の立案)、またパス図が不適切と考えられたときに何を加え、どの経路を変更するか（仮説の修正）などは、主観、直観によって行われる。たとえば重回帰における多重線形性に関連し、直観的には変数間に相関があるはずだが計算上は有意でないので解析を実行する、などというのは正しいやり方ではない。さらにデータを増やして検証するなどの手続きが必要であろう。方法は複雑になるほど、分析の具体的内容を把握しにくくなり、何かそこに独自の identity があるかのように錯覚しがちとなる。もしも、とりあえず使えるデータをできるだけ集めてきて重回帰式やパス解析のプログラムに input し、機械的に output させるというような「手法丸投げ」をすれば、たいていはわけのわからない結果に終わるだろう。無関係な変数であろうと、ケースー 1 だけ持ってくれば決定係数 1 が得られるという、重回帰の理論的帰結はその一例である。パソコンプログラムは生物学的真実とは無関係に、ただ与えられた計算を正確に実行するのである。

統計は人間の主観の客観化である。それはある場合は直観を明確にし、時にその修正を迫るものであるにしても、最終的な判断は人間の主観によって行われる。ある統計的手段が、人間の思考のどの部分を客観化しており、どの部分を人が自ら判断しなければならないかを知るには、理論的理解が不可欠となる。方法が複雑になるほどそれは困難だが、それをせずに使えば方法に振り回されることになるだろう。

引用文献

朝野熙彦 2000 入門多変量解析の実際　講談社サイエンティフィク
フラーリー・リードウィル 1990 多変量解析とその応用　現代数学社（田畑吉雄訳）
大垣俊一 1999　群集組成の多変量解析　Argonauta 1, 15-26

Ohgaki S, Takenouchi K, Hashimoto T, Nakai K 1999 Year-to-year changes in the rocky-shore malacofauna of Bansho Cape, central Japan: rising temperature and increasing abundance of southern species. Benthos Research, 54, 47-58
奥野忠一・久米均・芳賀敏郎・吉澤正 1971 多変量解析法　日科技連
塩谷實 1990 多変量解析概論　朝倉書店
Sipley, B 2000 Cause and correlation in Biology. Cambridge University Press
Sokal RR, Rohlf FJ 1981 Biometry 2nd ed. WH Freeman & Company
豊田秀樹・前田忠彦・柳井晴夫 1992 原因を探る統計学－共分散構造分析入門　講談社ブルーバックス
柳井晴夫・岩坪秀一 1976 複雑さに挑む科学－多変量解析入門　講談社ブルーバックス

（本文中では一々の引用を避けたが、以下のテキスト類も参考にした）
石居進 1975 生物統計学入門　培風舘
石村貞夫 1992 すぐわかる多変量解析　東京図書
木下栄蔵 1987 多変量解析法入門　啓学出版
丹慶勝一 2005 図解雑学多変量解析　ナツメ社
Zar JH 1999 Biostatistical Analysis, 4th ed. Prentice Hall