Argonauta 18: 17–30 (2010)

重回帰と偏相関、その後

大垣俊一

前報では重回帰、偏相関とパス解析について概略を紹介したが（大垣 2007）、基本的な内容に止まり、実際の論文等でデータ解析に使うにはカバーしきれていない部分がある点で不満が残った。しかしその後文献を読み足し、実際の適用場面でも若干考察を深めたので、それをもとにこのテーマについて再論したい。

私が重回帰的手法にこだわるのは、自分の研究テーマとの関係もある。生態学において要因分析、因果推論を行う場合の定番は、いわゆる操作実験である。実験区とコントロール区をランダムに配置し、問題とする要因のみを変更して経過を観察する。これは論理的に正確ではあるが、いつでも実行できるとは限らない。たとえば私が主なテーマとする生物相の長期変動では、変動要因として考えられるのは気候や海流といった大スケールで働く要因であって、これらをコントロールするような実験は事実上成立しない。とすれば変動する諸要因と生物側の反応について、できるだけ長い期間のデータをそろえ、両者の分散構造の中に各要因の寄与を推定するという行き方を選ばざるを得ない。つまり要因分析における操作実験と重回帰的手法はいわば車の両輪であって、どちらを欠いても生態学の内容は貧困なものになるだろう。

本稿では重回帰と偏相関について前報での不足を補いつつ、どのような場合にどの方法が適切かについて考えてみたい。なお記述に当っては、主に奥野ほか（1971）、Cohen et al.（2003）を参考にした。

重回帰

１．説明と予測

重回帰分析には説明と予測の二つの目的がある。テキスト類の多くはこの点について軽くふれて、あとは両者をまとめて解説しているが、実際の分析では異なる操作と解釈が必要なので区別する必要がある。「説明」は「分析」といってもよいと思うが、目的変数の変動に対し、それに関与するとみなされるいくつかの要因（説明要因）について、それぞれの寄与を評価する。「予測」は目的変数の変動を、適切な説明要因を選んでなるべく正確に言い当てようとする。生態学で重回帰を要因分析に使う場合、その用法はふつう「説明」に属する。たとえばある種の存在量に対し、A, B, C などの要因がそれぞれどの程度影響しているかを考える。それに対し「予測」は工学などで実行されることが多いだろう。ある製品を、X, Y, Z などの条件下で生産するときに、それぞれをどのような値にすると最も収量がよくなるかを考える、などである。ただ生態の研究でも、調査地点の水温が知りたいが限られた期間のデータしかなく、それ以外については近傍の観測値から推定しなければならない、といったケースがあ

る。このとき単に近隣の水温値から回帰したほうがよいか、気温など関連する要因を加えたほうが正確か、などは「予測」に属する問題になる。

英語の‘error’というタームに対し、残差と誤差という二つの訳語がある。残差とは、目的変数をいくつかの要因で説明しようとして、それらで説明し切れなかった残りの部分、というニュアンスがあり、これは「説明」と一体の概念である。これに対して「誤差」は、いくつかの要因を採用して目的変数を推定したが、いくらかはずれてしまったその程度、という意味だから「予測」のほうに対応する。この稿では主に「説明」を扱うので、error に対する訳語は「残差」とする。

２．残差をめぐる問題

重回帰を含む回帰分析では、残差は重要な意味を持つ。回帰係数の有意性を扱う前提として、残差にはのちの述べるようなさまざまな制限があるからである。ここで、このいわゆる回帰分析の条件は、説明変数のサンプル値そのものではなく、目的変数を説明変数で回帰したときの理論値からのズレ、つまり残差に対するものである。図 1 で残差分布は回帰線を水平に寝かせたときのサンプル値の、回帰線からのズレとして表現してある。この場合正規性その他の残差の要件は、図 1 左の③に適用されるのであって、目的変数分布（①）や説明変数分布（②）に対するものではない。従って残差分布の正規性が満たされないとき変換を用いるとしても、それは①や②ではなく③を正規分布に近づけるために行う。たとえば後段で紹介する類別変数の場合、説明変数の 0 や 1 の値が正規分布になるはずもないが、残差にはその可能性がある（図 1 右）。

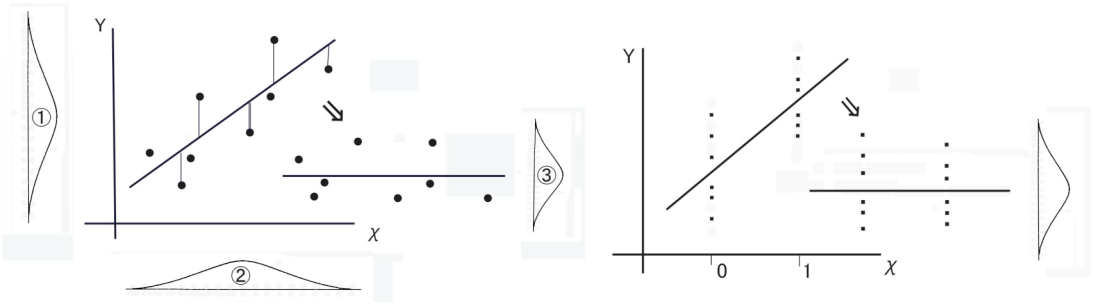

図 1. 回帰における残差分布のイメージ。X は説明変数、Y は目的変数。左）回帰直線に付随する線分が残差。右）説明変数が類別変数の場合の残差のイメージ。

重回帰分析における残差の要件は前報でもふれたが、ここでは奥野ほか（1981）に従って再提示する。

i) 残差は互いに独立である（独立性）
ii) 残差の期待値（平均値）は 0 である（不偏性）

iii) 残差の分散は互いに等しい（等分散性）
iv) 残差は正規分布に従う（正規性）

一つ目の独立性については、逆に独立でないとはどういうことかを考えたほうがわかりやすい。たとえば次の重回帰式、

$$Yi = \beta 0 + \beta 1 X1i + \beta 2 X2i + \cdots + \beta p Xpi + \varepsilon i$$

において、仮に説明変数 Xp を把握することができず、β pXpi + ε i の部分をまとめて残差 ε' i として扱ったと仮定する。するとこの ε' i は潜在変数 Xp が増加するほど大きな値をとることになり、各残差は互いにある種の従属関係、ないし関数関係にあることになる。つまり各 ε' i は独立ではない。したがってこのような場合は Xp を発見、分離して ε' i が互いに依存し合わない形にする必要がある。

二つ目の不偏性について。前報で解説したように、線形重回帰では最小 2 乗法に加えて、回帰線が X と Y の平均値を通過するように回帰式が定められる。このことから簡単な計算により、各残差の和は 0 となることが確かめられる。つまり残差不偏性は重回帰式の計算過程に組み込まれているから、プログラム計算をして回帰式を求める限り、特に気にする必要はないといえる。

三つ目の等分散性というのはわかりにくい表現である。残差は各 Xi の組に対して一つずつしかないので、それらの分散が互いに等しいというのは意味をなさない。しかしたとえば類別変数で回帰した場合のように（図 1 右）、同じ Xi の値に対して Yi が複数あるような状況を考えると、それぞれの Xi に対して残差の分散を考え、それらを比較することができる。一般的にいえば、横軸に各説明変数をとったとき、残差のばらつきがほぼ一定範囲に収まるということであろう。

四つ目の正規性は、図 1③の残差分布が正規分布に近いことを意味する。以上 i), ii), iii) は、偏回帰係数の有意性をめぐる t 検定に影響し、満たされないと各要因の寄与の評価が不正確になる。つまりこれらは「説明」的用法の要件であって、妥当な目的変数を推定することだけが目的の「予測」的用法においては問題にならない。ただ、このうち正規性については、一般に t 検定は、正規性からの逸脱に対して抵抗性があり（仮屋 1979, Underwood 1997）、重回帰の t 検定でも、標本数が大きければあまり問題にならないとされている（Cohen et al. 2003）。

実際の重回帰分析では目的変数の理論値や各説明変数に対して残差の分布を描き、それが一定範囲に収まるかどうか、また残差分布が正規分布に近いかどうかを調べる。重回帰分析の計算ソフトには、たいていそのための作図機能が備わっている。これらの図を見ながら、説明変数を加えたり、線形でなく 2 次曲線など非線形の回帰に変えるといった操作を行う。残差分析にはいろいろな側面があるが、ケース・バイケースであり、一般的には誤差のばらつきを描画して、ほぼ一定範囲に収まっていれば一応よしとするようである。

３．多重共線性

重回帰分析では、説明変数間に強い関係があると、説明、予測いずれにおいても支障をきたす。ここでいう強い関係には、大きく分けて次の２つがある。

i）ある説明変数が、他の説明変数の関数として定義される。たとえば $X_1 = k X_2$, $X_1 = X_2 + X_3$ など。

ii）関数的に定義できなくても、説明変数間に強い相関がある。

生態学の研究テーマでは、説明変数間の相関があるのがむしろ普通だから、重回帰分析を行おうとすると、必ずといってよいほど多重共線性の問題に行き当たる。多重共線性の不都合について、一つの説明は前報で述べた。要約すると、説明変数間の相関が高い場合、偏回帰係数の定義式の分母が 0 に近づく。そのため説明変数のわずかな変化によって偏回帰係数が大きく変動し、不安定になるというものである。

このほか次のような説明もある。偏回帰係数は、その定義式から、「偏回帰係数 b_2 は、X_2 から X_1 の影響を除いたあとでの、Y の X_2 に対する単回帰係数に等しい」（奥野ほか 1981）と解釈される。b_1 についても同様である。つまり、各説明変数に対する偏回帰係数には、あらかじめ他の説明変数の影響が含まれており、それを差し引いた値として示されている。このため、評価しようとする説明変数に強く相関するほかの説明変数があると、偏回帰係数はそれに影響されて著しく小さい（あるいは大きい）値を取り、目的変数に対する効果を正しく見積もることができなくなる。たとえば、説明変数間の相関が大きくなると、単回帰の回帰係数としてはプラスの大きい値を取っても、重回帰の偏回帰係数としては 0 に近い値や負になることもある。どちらが真かという問題はあるが、これでは少なくとも信頼できる分析にはならない。

多重共線性をどのように検出するかについては、符合の一致性、tolerance、VIF による方法などがある。符合の一致性とは、問題とする説明変数と目的変数の単回帰係数と、他の説明変数を加えた重回帰式での偏回帰係数を比較し、両者の符合が一致しない場合、無視できない多重共線性が発生しているとみなす。tolerance（許容値）は、「重回帰式における当の説明変数と他のすべての説明変数の重相関係数（R）を計算し、その２乗（R^2, 決定係数）を１から引いたもの：$tr = 1 - R^2$」と定義される。なお冗長度（redundancy）という用語があるが、これは目的変数に対する説明力が、各説明変数の間で重複している度合いのことである。その意味でこの R^2 に対応し、tolerance とは逆の概念となる。

R^2 が大きいことは、目的変数に対する当の説明変数と、他の説明変数群の間に強い連関があることを示している。従ってこれを１から引いた tr が小さいほど、多重共線性の影響を受けやすい。一般的には $tr > 0.1$ で多重共線性の影響は少ないと判断するが、この値やゆるすぎる（小さすぎる）という評価もある（Cohen et al. 2003）。

VIF は variance inflation factor の略で、$1 / (1 - R^2)$ で定義される。つまり tr の逆数で、$VIF > 10$（$tr = 0.1$）で多重共線性ありと判定する。variance inflation factor（分散拡大要因）というのは、これが数式的に「重回帰における目的変数分散／単回帰における目的変数分散」となり、説明変数間に相関がない場合に比べて、偏回帰係

数がどれほど大きくなるかを示す。つまり VIF は偏回帰係数の値の不安定性、ないし信頼度の低さの尺度である。

これらの方法で多重共線性が検出されたとき、重回帰の枠組みでこれに対処するには、相関する説明変数のいずれかを回帰式から除くしかない。その方法に、定番的なやり方はないようである。符合の一致性を使うときには不一致を示した説明変数を除き、ｔｒや VIF を用いるときには、低いｔｒや高い VIF を示す変数から順に除いてゆくのが常識的だ。しかしどうしても省けない、検討しなければならない要因というのもありうる。この場合はそれを残して他を省くことになるが、ここに何らかの恣意性が混入する危険もある。

４．偏回帰係数の検定

多重共線性の問題を何らかの方法でクリアーした後、各説明変数が目的変数の変動に対して効果を持つかどうかを判断するため、偏回帰係数の有意性を調べることになる。これは通常ｔ検定で行う。

$t\,cal = bi / \sqrt{\{Sf \times Se / (n-p-1)\}}$　（bi, Xi に対する偏回帰係数; Sf, 説明変数偏差平方和行列、逆行列対角要素; Se, 残差平方和；ｎ，サンプル数；ｐ，説明変数の数）

これによって算出したｔcal は自由度ｎ－ｐ－１のｔ分布に従い、有意性（bi = 0 からのへだたり）を検定することができる。この検定の要点は、次のように示される。「Xi 以外の説明変数はすべて用いるという条件の下で、なお Xi を加える意味があるか」（奥野ほか 1981）。上のｔcal の式を見ればわかるが、同じサンプル数であれば、説明変数が多いほどｎ－ｐ－１は小さく、分母全体は大きくなり、従ってｔcal は小さくなる。これは説明変数が多い場合、サンプル数もそれに応じて多くないと有意差が出にくくなることを意味している。

５．決定係数

重相関係数 R は、回帰時の目的変数の偏差平方和を S_R、ナマ（全体）の目的変数の偏差平方和を Syy として、$R^2 = S_R / Syy$　で表わされる。この式は、目的変数全体のばらつきの中で、回帰によって説明される割合を示しており、そのため R^2 は決定係数と呼ばれる。ここで注意すべきことは、決定係数は、それのみで要因（群）の関与の強さを保証するものではないということである。たとえば海岸線を南に下るほど海岸生物の種数が多くなるといったケースで、北から南に１から順に地点番号を割り振り、地点番号を説明変数、種数を目的変数として単回帰を計算すれば、強く有意な回帰係数と大きな決定係数が得られるだろう。しかしこのことから地点番号が種数をコントロールしているという結論はナンセンスである。つまり決定係数は、すべての説明変数が目的変数に因果的に連関している場合のみ、各説明変数の因果的寄与の度合いを示すのであって、そうでなければ生物学的意味を持たない単なる計算結果にすぎない。ただし逆に、R^2 が極めて低い値を取ったとすると、ほとんど関係のない説明変数群を選んでしまっていることはわかるから、むしろそのような否定的意味にお

いて決定係数の利用価値はあるといえる。

６．自由度調整済重相関係数

R^2 を上記の否定的目的に使うとしても、R は説明変数の数（ｐ）が大きいほど大きく、計算上サンプル数 p = n－1 において $S_R = S_{yy}$、つまり R = 1 となってしまうという性質がある。これは単回帰で n = 2 の場合を考えるとわかりやすい。単回帰でサンプル数が 2 ということは n = 2、 p = 1 であり、 p = n －1 が成り立っている。このとき平面上に 2 点しかないから、回帰直線はこの 2 点を通り、直線の上下にばらつく点はない。そのため y の回帰偏差平方和と全体の目的変数の偏差平方和は一致し、$R^2 = S_R / S_{yy} = 1$ となる。

しかし、無意味な説明変数であっても数を増やせばＲが自動的に 1 に近づくというのでは、決定係数としての R^2 の信頼性が損なわれる。そこで、ｎにくらべてｐが大きくなりすぎないよう、n－p－1 を少なくとも 10、なるべく 20 以上にすることが推奨されている。それができない場合、回帰の有効性を判定する指標として、自由度調整済重相関係数Ｒ*というものが考えられている。これはＲにおける偏差平方和をそれぞれの自由度で割り、分散で置きかえたもので、

$R^{*2} = V_R / V_{yy}$（V_R は回帰における目的変数分散、V_{yy} は全体の目的変数の分散）

と定義される。Ｒ*とＲの関係は、

$R^{*2} = \{(n-1)R^2 - p\} / (n - p - 1)$

右辺はｐ→大で分母→小、かつ分子→小だから、R^2 のように一方的に増加することはない。「自由度調整」というと先の矛盾を解消するために数式的操作を行っているような印象があるが、むしろ考え方の違いであり、結果としてＲ*は目的変数の数の影響を受けない形になっているというのが正確だろう。ただ、Ｒが「ばらつきに対するばらつきの割合」として回帰による説明程度をストレートに表現するのに対し、Ｒ*は「分散に対する分散の割合」として、具体的に意味するところがあいまいになっている。決定係数として使うのであれば、通常の重相関係数Ｒの方がわかりやすい。

７．類別変数の組み込み：ダミー変数

重回帰分析では、説明変数として連続変数のほか、類別的な変数を含めたい場合がある。たとえば海岸生物の存在量に対する要因として、水温、塩分などは連続変数だが、基盤の地質が堆積岩か火成岩かなどは類別的な変数である。これらを一括して評価するために、類別変数に対してダミー変数を割り当てるやり方がある（表１）。岩質の例でいえば、元データ表の列項目に堆積岩、火成岩、その他、のように変数を並べ、ある地点が堆積岩海岸であれば、その行に、1, 0, 0 のように数字を割り当てる。同様に火成岩なら 0, 1, 0、その他なら 0, 0, 1 である。ダミー変数は 0 や 1 でなくても、0, 10 でも 1, 2 であっても、偏回帰係数の有意検定では、結局標準化されるから

結果に影響しない。普通は最もわかりやすい 0, 1 を用いる。このような 2 進法的変数を他の連続変数と共に重回帰の枠組みに取り込むのは直観的に違和感もあるが、理論的に可能であり、また普通に行われている。たとえば堆積岩かそれ以外かという 2 変数を考えると、回帰のイメージは先に示した図 1 右のようになる。

類別変数を取り込むときに注意すべきことは、タイプ分けしたすべての変数を分析に用いることはできないということである。先の例では、3 つの岩質タイプをすべて用いると、堆積岩、火成岩でなければ必ずその他、火成岩、その他でなければ必ず堆積岩、というような関係が成り立っている。これは多重共線性の項で述べた禁止事項、「ある説明変数が、他の変数の関数として定義される」に該当する。そのためそのまま計算プログラムを実行すると、「計算不能」の警告が表示される。こういう場合は、どれか一つの変数（その他、など）を省かざるを得ない。カテゴリーが 2 つなら、表の一つの列の中でその条件が、「あり」なら 1、「なし」なら 0 を割り振れば計算できる。

偏相関

1．ノンパラメトリクス偏相関

パラメトリクスの相関（Pearson 積率相関）に対して Spearman（相関係数 ρ）や Kendall（相関係数 τ）の順位相関があるように、パラメトリクスの編相関係数に対しても、Spearman 型や Kendall 型の偏相関係数が考えられている。これらは 3 変数の場合なら、通常の偏相関係数の式

$$r_{12\cdot3} = (r_{12} - r_{13} r_{23}) / (1 - r_{13}^2)(1 - r_{23}^2)$$

（$r_{12\cdot3}$ は 3 の影響を除いた $\chi 1$ と $\chi 2$ の偏相関係数、r_{12} は $\chi 1$ と $\chi 2$ の単相関係数）

の r のところに、Spearman 係数 ρ か Kendall 係数 τ を入れて求めることができる（Conover 1999）。この程度ならのちの t 検定まで含め、もとになる順位相関係数がわかっていれば手計算でも求まるし、4 変数以上でも、Spearman 型ならパラメトリクス型の偏相関プログラムに ρ や τ を代入して計算するやり方がある。またあまり一般的ではないが、順位偏相関のプログラムそのものを搭載しているソフトもあるようだ。

パラメトリクス偏相関の検定も t 検定で行うことができる。3 変数なら

$$t_p = r_{y1\cdot2} / \sqrt{\{(1 - r_{y12}^2) / (n - 3)\}}$$

（$r_{y1\cdot2}$ は $\chi 2$ の影響を除いた $\chi 1$ と y の偏相関係数）

これが自由度 n－3 の t 分布に従うことをもとに検定する。Spearman 型では、t 式の偏相関係数のところに、ρ から求めた先の偏相関係数を代入して t を計算する。τ については母集団分布が明らかでないため有意性の検定はできないらしい。

パラメトリクス型の偏相関では、有意性の検定に当り、各変数母集団の多変量正規分布が前提になる。ノンパラならばそういう窮屈な前提は必要ないだろうと期待するわけだが、残念ながらこの場合も distribution free とはいかない。その有意検定は多変量分布（multivariate distribution）に影響される（Conover 1999）。多変量分布関数というのは、たとえば多変量正規分布のようなものをいうのだろうが、この場合に具体的に何なのかは詳しく書いていないのでわからない。ちなみにノンパラメトリクスのU検定では、比較する２母集団の「分布形相等」が検定の条件であるとされている（石居 1975, Underwood 1997）。一方ｔ検定では、先に回帰のところで見たように、母集団が正規分布からはずれても検定への影響は少ないとするのが一般的である。とすると、パラメトリクス型を避けてノンパラ型を採用する意味はあまりないことになる。ノンパラメトリクス偏相関が、実際の研究にも市販のソフト類でもあまり採用されていないのは、そのためかもしれない。

２．半偏相関（Semi-partial correlation）

生態学の論文では、要因分析に当って半偏相関という手法が使われることがある（たとえば Lagos et al. 2008）。偏相関の場合、ｙ, χ1, χ2, … χｎ の変数があって、この中のｙとχ1の関係が知りたいときは、ｙをχ2, … χｎで回帰した残差と、χ1をχ2, … χｎで回帰した残差を求め、これらの相関を計算する。これに対し半偏相関は、χ1だけをχ2, … χｎで回帰して残差を求め、これとナマのｙ値の相関を計算する。つまり、偏相関ではχ2, … χｎの影響を除いたｙとχ1の相関を調べ、ｙとχ1の関係は対等で双方向的なのに対し、半偏相関ではχ2, … χｎの影響はχ1だけから除かれることから、ｙとχ1の関係に非対称性がある（図２）。

図２．偏相関（左）と半偏相関（右）のイメージ。

私の見た限り、日本のテキストや論文では半偏相関はほとんど扱われていないが、欧米では要因分析における各変数の "true effect" ないし "unique contribution" を検出する方法として推奨されている（Tabachnick & Fidell 1983, Freckleton 2002）。半偏相関係数は、４変数以上になると複雑なので、プログラムで求めるしかないが、ソフトによっては係数のみ表示して、有意検定まで行わないことがある。その場合は、求めた相関係数を次の式に代入してｔ値を求めることができる。次のｔｓは自由度ｎ－ｑ－２のｔ分布に従う。なお、偏相関の検定も同じ式により、ｒxyのところに偏相関係数を代入すればよい。

$$t_s = r_{(xy)q} / \sqrt{\{(1 - r_{(xy)q}^2) / (n - q - 2)\}} \quad \cdots ①$$

（r (xy)q は、x に対してのみ他の q 個の変数を固定したときの、x と y の半偏相関係数）

ここで偏相関と半偏相関を、3 変数の場合について比較してみる。

偏相関係数：$r_{y1\cdot 2} = (r_{y1} - r_{y2}\ r_{12}) / \sqrt{\{(1 - r_{2y}^2)(1 - r_{12}^2)\}}$　（再掲）
（$r_{y1\cdot 2}$ は χ_2 の影響を除いた y と χ_1 の偏相関係数、r_{y1} は y と χ_1 の単相関係数）
半偏相関係数：$r_{y(1\cdot 2)} = (r_{y1} - r_{y2}\ r_{12}) / \sqrt{(1 - r_{12}^2)}$
（$r_{y(1\cdot 2)}$ は y と、χ_2 の影響を除いた χ_1 との半偏相関係数）

これらを見ると分子は同じだが、分母は偏相関にあった$(1 - r_{2y}^2)$ が半偏相関では欠けている。ここで $0 < r_{2y}^2 < 1$ だから、$(1 - r_{2y}^2)$ は 0～1。これで割っているから、$r_{y1\cdot 2} > r_{y(1\cdot 2)}$。つまり同じデータについて計算すると、偏相関係数は半偏相関係数よりも必ず大きい。このことは検定に影響する。偏相関係数も半偏相関係数も、共に①式で t を求めて検定する。両者の違いは、r (xy) のところに偏相関係数を入れるか半偏相関係数を入れるかだけである。そのため半偏相関のほうが有意差が出にくくなる。このことは実際に計算してみると実感されることで、10 程度の変数を扱おうとすると、半偏相関では有意差を検出するのに数百のサンプル数を必要とする場合がある。

偏相関に対して半偏相関を用いるメリットについては次のようにに考えられる。まず半偏相関の 2 乗（sr^2）は、理論的に目的変数の分散に対する、各説明変数の寄与を表わす。正確には「その説明変数を加えることによる、重回帰式の説明率の増加分」を示している（Cohen et al. 2003）。これが、sr^2 が "unique contribution"と呼ばれるゆえんである。そもそも y の値そのものに対する χ_1 の影響を知りたいのであれば、y の残差でなくナマの値を使う方がわかりやすい。また処理の非対称性によって $\chi_1 \rightarrow$ y の説明という性格が加わったとみることもできるから、その意味では半偏相関は重回帰に近いといえよう。

分析の実例

ここで仮に構成したデータ（次ページ表 1）をもとに、類別変数を含む重回帰、偏相関、半偏相関の分析を実際に行ってみる。これは人工データだが、一応、ある地域の海岸 20 地点で生物相を調べ、その中の南方性の種の存在割合（S %）を計算したと想定している。これに各地点の冬の水温と塩分、基盤の岩質（A, B, その他）を対応させた。また泥岩を多く含む海岸かそうでないかの情報も加えてある。岩質以下は類別変数だから、それぞれ該当は 1、非該当は 0 のダミー変数を割り当てた。目的変数は S %、説明変数は連続型（間隔変数）２、類別型 4 の計 6 変数である。

地点	S %	水温	塩分	岩質	岩質	岩質	泥岩
1	5.7	10.3	32.9	1	0	0	1
2	13.4	10.9	33.0	0	0	1	0
3	2.9	11.2	33.2	0	0	1	0
4	17.2	13.8	34.1	0	1	0	0
5	20.1	13.7	34.1	0	1	0	0
6	30.5	14.0	34.0	0	1	0	1
7	34.8	15.2	34.6	0	1	0	1
8	31.0	15.3	34.4	0	1	0	1
9	26.1	16.2	34.5	0	1	0	1
10	37.1	15.9	34.5	1	0	0	0
11	39.4	17.5	34.8	1	0	0	1
12	35.1	17.8	34.0	1	0	0	1
13	44.9	18.0	33.3	0	1	0	0
14	40.6	17.3	34.7	0	1	0	0
15	40.1	17.6	34.7	0	1	0	1
16	52.7	17.8	34.7	0	0	1	0
17	34.5	17.7	34.7	0	0	1	0
18	10.5	17.3	34.7	1	0	0	1
19	16.5	16.7	34.7	1	0	0	1
20	17.6	16.0	34.6	1	0	0	1

表１．サンプルデータ。水温は℃、塩分は psu（‰）単位。「岩質他」は A, B 以外の岩質。

まず重回帰分析を行う。表を見ると、岩質の３項目間に、「A, B でなければかならずその他」のように、明らかな共線性が存在することがわかる。また泥岩は「岩質他」と全く重なっておらず、両者の間に強い逆相関がある。このままプログラムを走らせても「実行不能」となることは明らかだから、ここで「岩質他」を検討から除き、残り 5 項目に絞る。この時点で、「その他の岩質」の、S%への寄与を見積もることは断念しなければならない。

重回帰分析に入って、まず残差を検討する。結果は図３のようになる。

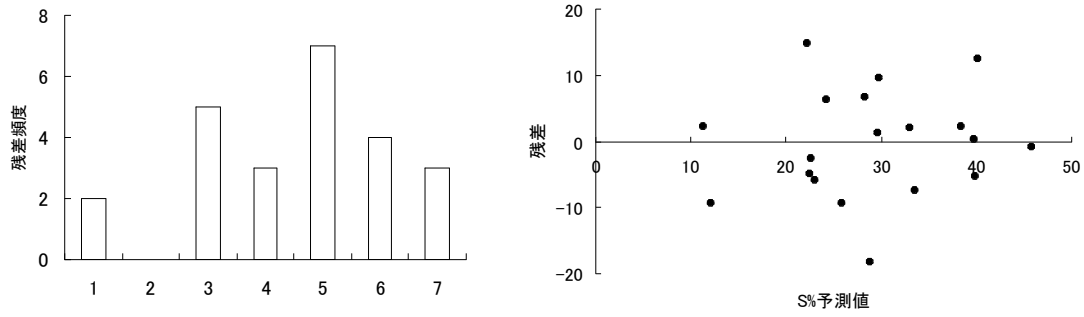

図３．表１にもとづく重回帰の残差分析。左図横軸の 1–7 は、残差 −20〜+20 までを 5 ずつに区切った階級を示す。

左図の正規性はほぼ満足されているように見える。右図の目的変数に対する残差分布も、問題があるようにはみえない。各説明変数についての残差分布も見たほうがよいが、煩雑になるのでこのまま先に進める。5 つの説明変数の標準偏回帰係数と、t 検定の結果としての P 値、tolerance は表 2 のようになる。

説明変数	標準偏回帰係数	P 値	tolerance
水温	0.849	0.004	0.406
塩分	–0.106	0.679	0.390
岩質 A	–0.237	0.394	0.336
岩質 B	0.083	0.733	0.425
泥岩	–0.098	0.637	0.589

表 2. 表 1 に基く重回帰分析の結果。有意な P 値に下線を付す。

tolerance はすべての変数について 0.3 を越えており、多重共線性の問題はないとみてよい。有意な P 値は水温のみに見られ、回帰係数の符号から、冬季水温が高い地点ほど南方性種の割合が高いことがわかる。決定係数 R^2 は 0.66 で、この 5 変数で目的変数分散の 2/3 を説明する。

重回帰を実行すると、たいてい偏相関も同時に計算される。そこで表示された偏相関係数、半偏相関係数と自由度をもとに t 検定を行い、P 値を求めた結果は表 3 の通りである。

説明変数	偏相関係数	P	半偏相関係数	P
水温	0.680	< 0.001	0.541	< 0.02
塩分	–0.112	ns	–0.066	ns
岩質 A	–0.229	ns	–0.137	ns
岩質 B	0.093	ns	0.054	ns
泥岩	–0.128	ns	–0.075	ns

表 3. 偏相関分析の結果。

重回帰の結果と同じく、水温のみに強い相関が認められた。この表を見ると、先に示したように半偏相関係数は偏相関係数より必ず小さくなり、同じデータに対して有意性を検出しにくくなっていることが確認される。

方法間の比較

単回帰・単相関：　複数の説明要因があるとき、その相互関係は不問として、各説明変数と目的変数との単相関とを個別に計算するというやり方は、素朴だが場合によっては有効である。サンプル数は数十程度で説明変数が十くらいあり、そのうちいくつ

かは相互連関しているという場合に重回帰を使うと、多重共線性を避けるために説明要因を減らさざるを得ず、また偏相関や半偏相関では説明変数に対するサンプル数が少なく、有意性が出にくいという難点がある。しかし単相関を使えば、要因を減らさずかつ有意差をより高い頻度で検出することができる。もちろん、有意性を検出することが研究の目的ではないが、それが関与の可能性の強い指標を選択する際の客観的な基準を与えるものである以上、すべての説明要因が非有意だったという結果は好ましいものではない。

一方で単回帰、単相関は、他の条件がすべて一定とみなした上で、算出、検定するから、その結論は暫定的なものにならざるをえない。単回帰・単相関は簡便であり有意な結果を得やすい。しかしそれだけで確定的な結論を導くのは多くの場合無理であり、これらで要因を絞った上、操作実験や、よりくわしい観察などの外部情報によって決着をつけることになるだろう。

重回帰・偏相関：　前報で重回帰は未知要因を残差の中に含めて考えられるのが生態学の分析ではメリットになると書いたが、この考え方は統計学的には不可とされるだろう。残差の項で述べたように、目的変数に影響を与える未知の要因がある場合は残差の要件を満たさず、それを加えた回帰式を再構成しなければ正しい分析はできない。しかし現場には現場の論理がある。実際の分析では、目的変数に影響するすべての要因を知ることができるとは限らない。野外データにおいては特にそうである。また存在することが推測できても、測定できないなどでデータが手に入らないこともある。こうした場合でも、未知要因が残差の中に含まれているならば、すでに取り込まれている要因の偏回帰係数の説明率は小さくなるはずだから、定性的評価としては有効でありうる。

それでも残差に未知要因を含むことを前提に分析するのは許されないというのであれば、重回帰分析の実用上のメリットは著しく損なわれる。もともと重回帰には残差をめぐる制限に加え、多重共線性という深刻な問題がある。説明変数間の連関は生態学ではむしろ常態で、多重共線性を避けようとすると、検討したい要因を除去しなければならないこともある。また、要因除去に当って恣意性が混入する危険もある。たとえばＡとＢが連関していてどちらかを除去しなければならないとき、Ａを残したほうが自らの筋立てに有利であると思えば、分析者はＢを除去したくなるかもしれない。

一方偏相関、半偏相関は、目的変数に影響を与える要因がいくつかに限定され、かつ説明変数間に強い連関があるときに向いている。要因は選ばれたものに限定されていることが分析の前提だが、そうでなければやってはならないというわけでもない。単相関よりも多くの要因を同時に検討できることに満足して、その枠組みの中で暫定的な結論を導き、あとの検証を外部情報にゆだねる、という考え方もありうる。半偏相関は偏相関に比べ、説明要因→目的要因の方向性を持ち、その２乗値が目的変数分散に対する寄与率を示すことから、要因分析に当っては偏相関よりも理論的に妥当と考えられる。偏相関より有意性が出にくいが、説明変数が多くなく十分なサンプル数が得られるなら、こちらを使うべきだろう。偏相関には多変量正規分布の制限がある

が、正規分布の要件は特にサンプル数が多いときには深刻でないとされているので、それほど気にする必要はないのかもしれない。

パス解析：　重回帰、偏相関の発展形態としてのパス解析については前報である程度紹介した。これはある意味正確だが、多くの変数についてのデータを必要とし、理論が複雑かつ計算量も多く、大がかりなものとなる。このことから、あるテーマの中の要因分析的な部分にだけ適用するには使いにくい。加えて結論は良くいえば包括的、悪くいうとあいまいなものになりがちで、こうした特徴が、生態学の分析においてパス解析がほとんど使われていないことの理由かもしれない。しかし心理学や社会学では多くの適用例があるようであり、生態学においても全体構造をメインテーマとして採用する、というような方向性はありうるかもしれない。

以上、各方法の特徴について、図４にまとめた。

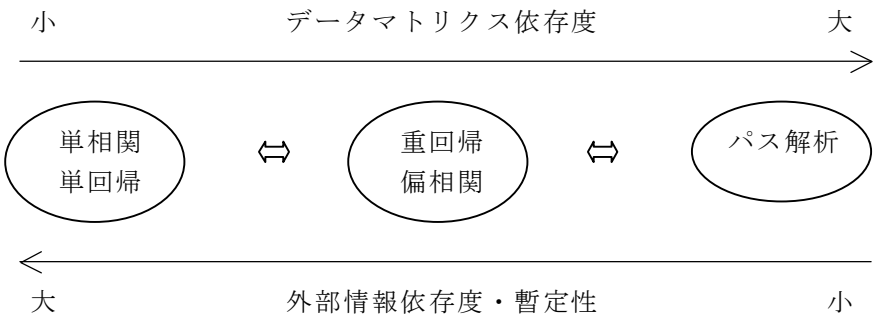

図４．回帰・相関系分析のまとめ。

単回帰、単相関は分析の暫定性が強く、それだけ実験やより詳細な観察などの外部情報への依存度が高い。逆にパス解析はデータマトリクス（たとえば表１）への依存度が高く、その枠内でできる限りの結論を得ようとする。重回帰と偏相関はそれらの中間的性格を持つ。

最後に言いたいことは、「単純＝劣っている」という単純な考え方はナンセンスだということである。複雑なやり方には同時に固有の制限やあいまいさが伴い、複雑化させるほど一方的に分析精度が上がるという保証はない。ここで述べてきたような各手法の特徴をふまえ、テーマに合った方法を選択することで、要因分析をより適切な形で生態学に取り入れることができると考えられる。

引用文献

Cohen J, Cohen P, West SG, Aiken LS (2003) Applied multiple regression / correlation analysis for the behavioral sciences, 3rd ed. Lawrence Erlbaum

Associates Publishers
Conover WJ (1999) Practical nonparametric statistics, 3rd ed. John Wiley & Sons
Freckleton RP (2002) On the misuse of residuals in ecology: regression of residuals vs. multiple regression. J Anim Ecol, 71, 542–545
石居進（1975）生物統計学入門. 培風館
刈屋武昭（1979）回帰分析の理論. 岩波書店
Lagos NA, Castilla JC, Broitman BR (2008) Spatial environmental correlates of intertidal recruitment: a test using barnacles in northern Chile. Ecol Monogr, 245–261
大垣俊一（2007）重回帰、偏相関、パス解析. Argonauta, 13, 3–23
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
Tabachnik BG & Fidell LS (1983) Using multivariate statistics. Harper & Row
Underwood AJ (1997) Experiments in ecology. Cambridge University Press