データから因果を推定する

清水顕史

１．相関係数の問題点

２変数間の関連をみる指標としては（ピアソンの積率）相関係数（r）がよく用いられる。rは２変数間の直線的関係（一次の線形性）のみを示すことができる尺度である。相関という言葉は、まるで２変数間の様々な因果関係を類推できる指標のように誤用されることが多い。相関係数を鵜呑みにすることで起こる多くの誤りの中でも、（生物）研究上特に重要なものとしては、集団の階層化(stratification)問題がある。

右の散布図で示した２変数A,Bの間にはr=−.641の1%水準で有意な負の相関がみられる。例えば変数A,Bがそれぞれ遺伝子A,Bの発現量を表しているとすると、有意なrはA,B間にある負の制御（片方の発現が増えるともう片方が減る、もしくは片方が減るともう片方が増える。上流か下流かなどカスケードの位置関係は分からない）を示しているようにみえる。

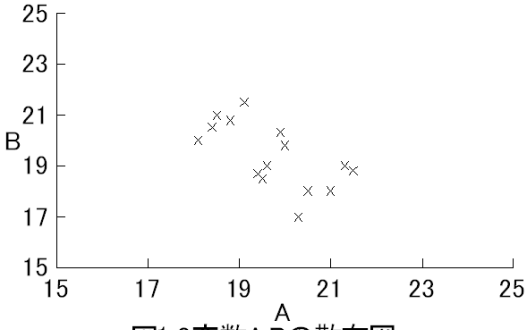


図1 2変数A,Bの散布図

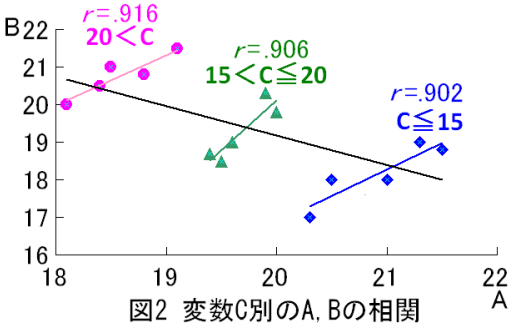


図2 変数C別のA, Bの相関

しかし、調査した15個のサンプルについて、遺伝子Cの発現量別にA,Bの発現量を調べてみると左図のようになったとする。つまり、AとBの相関係数rは、Cの発現量が20＜の時は.916であり、15＜C≦20の時に.906、C≦15の時に.902と、強い正の相関が見られた。これはA,B間には強い正の制御があることを示している。

図１のように２変数間のみで見た関連(association)と図２のように他の変数による条件付き関連との方向が逆転してしまう現象は、シンプソンのパラドクスと呼ばれている。この場合、Cによる影響に気付かずにA,Bのみのデータだけから解釈すれば誤った結論を導き出してしまうことに注意しよう。AとBのみの場合に検出された負の相関は、偽の(Spurious)関連もしくは周辺(Marginal)関連と呼ばれるものである。

2. 層別

2 変数 A,B 間の相関関係を別の変数で区別するデータ解析手法を層別(stratification)という。層別は、気付かなければ見落としてしまうような条件付関連を検出する高度な解析技法である。ただし図 2 のような A,B 間の条件付関連を見つけるためには、C のデータがなければはじまらない。このことは、解析に用いるデータとして何を取得すべきかの取捨選択の重要性を意味している、もしくは遺伝子発現解析のような場合ならデータ探索範囲の網羅性が如何に重要かを教えてくれる。

先頁の例の場合、A,B 間が周辺関連でしかないことは、有意ではあるが値がそれほど高くない相関係数から見積もることができるかもしれない（それに気づく洞察力がある場合)。相関係数が有意であることは、一次直線関係が無相関（独立）であることのみを示しているため、それが直接的な関連であるか周辺関連でしかないかを教えてくれない。相関係数から変数間の関連を見積もる場合は、$r = 0$ という帰無仮説が棄却されるかどうかだけでなく、r の絶対値の大きさ(0.8 か 0.7 以上を strong correlation として区別することが多い)や信頼区間も判断材料にするべきである。ちなみに相関係数 r の二乗(R^2)は決定係数と呼ばれ、r を計算する際に想定する直線が 2 変数の散らばりをどれだけよく説明しているかを表す数値である（例えば $r = 0.8$ とは、散布図の 64%が直線で説明できることを意味する)。

多変量データの中の 2 変数間の関連を見る際、周辺関連に騙されないようにするためには、以下に示す偏相関係数を用いるのがよいだろう。

3. 相関係数と偏相関係数

2 変数の(1 次の)相関を表す指標として(単純)相関係数 r がある。変数対(a_i, b_i)、 i =1～n の r は、

$$r = \frac{\sum a_i b_i - \dfrac{\sum a_i \sum b_i}{n}}{\sqrt{\left(\sum a_i^2 - \dfrac{\left(\sum a_i\right)^2}{n}\right)\left(\sum b_i^2 - \dfrac{\left(\sum b_i\right)^2}{n}\right)}} \qquad \text{式(1)}$$

を計算すればよい。ちなみに Σx_i は $\sum_{i=1}^{n} x_i$ の意味である。

3 変数 a, b, c がある場合の 2 変数 a, b の偏相関係数は、$r_{ab \bullet c}$ と表せる。変数 c と変数 a および b との交互作用を取り除くことで、$r_{ab \bullet c}$ は変数 c によって値が変化しない。$r_{ab \bullet c}$ は、a と b、a と c、b と c の単純相関係数 r_{ab}, r_{ac}, r_{bc} から下式のようにして計算できる。

$$r_{ab \bullet c} = \frac{r_{ab} - r_{ac} r_{bc}}{\sqrt{\left(1 - r_{ac}^2\right)\left(1 - r_{bc}^2\right)}} \qquad \text{式(2)}$$

下の例データで、

i	a_i	b_i	c_i
1	18.1	20.0	11.0
2	18.4	19.0	13.0
3	15.0	21.0	23.6
4	14.0	22.0	21.0
5	19.1	21.5	24.8
6	18.0	18.5	17.2
7	17.0	18.7	16.8
8	16.0	19.0	17.0
9	19.9	18.0	12.2
10	20.0	19.8	13.0
11	18.0	18.2	13.0
12	17.0	18.5	13.0
13	21.0	16.5	8.0
14	24.0	17.0	2.3
15	21.5	18.8	7.0

式(1)より相関係数$r_{ab} = -.636$、$r_{ac} = -.780$、$r_{bc} = .786$、$r_{ac} = .786$が計算でき、式（２）より偏相間係数$r_{ab\bullet c} = -.059$、$r_{bc\bullet a} = .600$、$r_{ac\bullet b} = -.588$が計算できる。標本数が 15 個の場合、相関係数は.514 以上で 5%有意であり、.641 以上で 1%有意、.760 以上で 0.1%有意となる。標本数を n 個、固定する変数を q 個、変数 a と b の偏相関係数 $r_{ab\bullet}$ とするとき、偏相間係数の検定は以下の検定量を計算する。

$$t_0 = \frac{|r_{ab\bullet}|\sqrt{n-q-2}}{\sqrt{1-r_{ab\bullet}^{\ 2}}}$$

t_0 は自由度が n-q-2 の t 分布に従うとし、棄却されるとき、偏相間係数は 0 でないとみなす。上のデータの場合、n=15, q=1 なので偏相間係数が.532 以上の時 5%有意となる。

相関行列や偏相関行列は、R を使うと簡単に計算できる。以下では、入力するコマンドを赤字で示す(入力ファイル。**作業用ディレクトリを指定しておく(講義用 Web ページから Cortest.txt をダウンローしたフォルダを指定する**)。

```
X <- as.matrix(read.table("Cortest.txt"))      #多変量データの読み込み X とする
Y <- cor(X)                                    #相関行列の計算結果を Y とする
Y                                              #Y を表示
invY <- solve(Y)                               #Y の逆行列を invY とする
O <- diag(1/sqrt(diag(invY))) #O:invY の対角成分の平方根の逆数の対角行列
P <- -O %*% invY %*% O                         #偏相関係数行列を P とする
P                                              #P を表示
```

4. 共分散選択

偏相間係数が 0 であるとは、対象とする 2 変数間が（残りの変数の影響を取り除く条件付の場合に）独立である（無関係）ことを意味する。偏相間係数の値が小さい変数間について、その変数間が条件付独立であると仮定した関連構造モデルを採用する方法を共分散選択という。

先の 3 変数の場合の例では、$r_{ab\bullet c}$ は 5%で有意でなく値も小さい。そこで共分散選択では $r_{ab\bullet c}$ を 0 という制約のもとで偏相関係数 $r_{bc\bullet a}$ と $r_{ac\bullet b}$ を計算しなおす。r_{ac} と r_{bc} は制約の影響を受けないので、

$$r_{ab\bullet c} = 0 = \frac{r'_{ab} - r_{ac}r_{bc}}{\sqrt{\left(1-r_{ac}^2\right)\left(1-r_{bc}^2\right)}}$$

より $r'_{ab} = -.613$ と計算できる。これを元に $r_{bc\bullet a}$ と $r_{ac\bullet b}$ を再計算すると、$r_{bc\bullet a} = .622$、$r_{ac\bullet b} = -.611$ となる。制約をいれることによって変化する相関行列、偏相間行列をまとめると、

	制約なし	$r_{ab\cdot c} = 0$
相関行列	$\begin{pmatrix} 1 & & \\ -0.636 & 1 & \\ -0.780 & 0.786 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ -0.613 & 1 & \\ -0.780 & 0.786 & 1 \end{pmatrix}$
偏相関行列	$\begin{pmatrix} - & & \\ -0.059 & - & \\ -0.588 & 0.600 & - \end{pmatrix}$	$\begin{pmatrix} - & & \\ 0.000 & - & \\ -0.611 & 0.622 & - \end{pmatrix}$

となる（対象行列なので下三角のみで示している）。

$r_{ab\bullet c} = 0$ の制約を受け入れるかどうかの判断基準として、共分散選択では逸脱度という指標を用いる。標本数 n とし、制約なしモデルと制約ありモデルの分散共分散行列をそれぞれ S_0、S_1 とするとき、逸脱度 d は

$$d = n \ln(|S_1|/|S_0|)$$

で計算できる。相関行列 $\boldsymbol{R}$ と分散共分散行列 $\boldsymbol{S}$ には $\boldsymbol{R} = \boldsymbol{DSD}$ という関係がある。ちなみに $\boldsymbol{D}$ は $\boldsymbol{S}$ の対角成分の平方根の逆数を対角成分として持つ対角行列である。それぞれのモデルの分散共分散行列(対象行列なので下三角のみで記す)は

制約なし

$$\begin{pmatrix} 6.265 & & \\ -2.370 & 2.217 & \\ -11.526 & 6.905 & 34.830 \end{pmatrix}$$

$r_{ab\cdot c} = 0$

$$\begin{pmatrix} 6.265 & & \\ -2.285 & 2.217 & \\ -11.526 & 6.905 & 34.830 \end{pmatrix}$$

となり、逸脱度 d は 0.052 となる。一つの制約を加えたモデル間の逸脱度 d は自由度 1 のカイ 2 乗分布に従うことが分かっている。d=0.052 となる p 値は 0.819 であり、これは $r_{ab\bullet c} = 0$ の制約モデルが正しいとき d が 0.052 以上になる確率は.819 以上であることを意味する。従ってこの制約モデルを受け入れてよいことになる。$r_{bc\bullet a} = 0$ の制約モデルでの d は 6.69 (p=0.010)、$r_{ac\bullet b} = 0$ の制約モデルでは d=6.36(p=0.012)でこれらの制約モデルは受け入れられない。

共分散選択による偏相間係数の取捨は、より単純な関連構造モデルを提示してくれる。得られた関連構造モデルはグラフで表現されるので「グラフィカルモデリング」と呼ばれ、より直接的な変数間の関連を見出す手法として利用されている。

5. グラフィカルモデリング

グラフィカルモデリングによる多変量変数の解析例を記す。データ例は、リン欠乏ストレス条件で育成した 55 系統のイネの 6 種類の形質データを解析したものである(Shimizu et al. *Theor Appl Genet* 117:987-996)。グラフィカルモデリングには MIM というソフトウェアを使用した(http://www.hypergraph.dk/)。

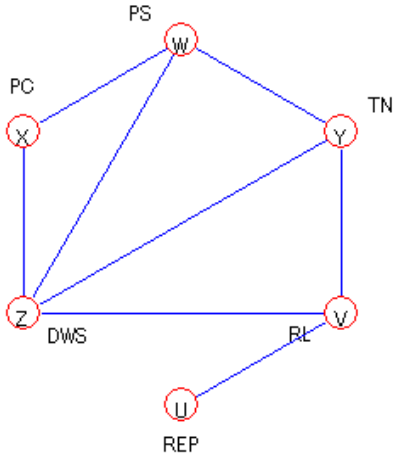
0) 入力用データを準備する。MIM のフォーマットに合わせた入力データの一部を下に示す。"cont"は 6 種類の変数(U～Z)が全て連続変量であることを意味し、"label"は U が"REP"、V が"RL"、W が"PS(地上部のリン含有率)"、X が"PC"、Y が"TN"、Z が"DWS"というそれぞれの形質名に対応させるコマンドである。"read UVWXYZ"以下の各行に個体毎のデータをタブ区切りで記入する(順番は label で指定した U～Z の順)、データの最後尾は"!"で区切られる。"%"はコメント行で入力データとしては読み込まれない部分である。ファイルの拡張子は dat とする（くわしくは MIM の help を参照)。

```
% CSSL Pdef
cont UVWXYZ
label U "REP" V "RL" W "PS" X "PC" Y "TN" Z "DWS"
read UVWXYZ
1.14    14.2    13.7    9.53    1       0.689
1.02    12.9    17.1    11.42   1.11    0.667
0.98    13.8    16.8    10.24   1       0.608
  ⋮       ⋮       ⋮       ⋮       ⋮       ⋮
1.27    13.9    17.3    8.31    1       0.48
1.26    17.1    17.6    11.92   1.4     0.677
1.24    17.6    15.8    12.85   1.44    0.81025
1.2     15.1    12.1    8.75    1       0.716
!
```

1) MIM を起動して入力データを読み込む。メニューバーの[File]をクリックし、出てきたタブから[Input…]をクリックし、ファイル検索で入力用データを選ぶ。データが正常に読み込まれたら"Reading completed"と表示される（エラーの場合は入力ファイルのフォーマットが正常であるかを再度確認する)。
2) 初期モデルとして完全な関連構造モデルを選択する。メニューバーの[Model]をクリックし、[Saturated Model]を選ぶ。この状態で、メニューバーの[Fit]をクリックしタブから[Fit]を選択すると、全ての 2 変数間の関連を仮定する完全な関連構造モデルを選択することになる。この状態でメニューバーの[Fit]をクリックし[Show estimate]を選び、出てきたウィンドウの"… covariance"、"…correlations"、"…partial correlations"をチェックするとそれぞれ、完全な関連構造モデルにおける共分散行列、相関行列、偏相関行列を表示できる。
3) 共分散選択法による、より適したモデルの選抜を行う。メニューバーの[Select]をクリックし、[Stepwise]を選ぶ。デフォルト設定で[OK]を押すと、共分散選択による最適モ

デルが自動で選抜される。多変量データにおける共分散選択において、完全な関連構造モデル(M0)の偏相間係数の一番小さいものをゼロと置く制約モデル(M1)との逸脱度を考え、M0 と M1 との逸脱度が自由度 1 のカイ 2 乗分布において有意確率 p>0.5 の場合に制約モデル(M1)を受け入れる。次に受け入れられた制約モデル(M1)の偏相間係数のうち一番小さいものをゼロと置く新たな制約モデル(M2)を考え、M1 と M2 の逸脱度を検定する。M(t)モデルの最小の偏相間係数をゼロと置く制約を加えたモデル M(t+1)との逸脱度が自由度 1 のカイ 2 乗分布において有意確率 p>0.5 であり続ける限り制約モデルを選択する。

メニューバーの[Graphics]から[Graph]を選ぶと選抜された関連構造モデルの無向グラフ（右図）を表示することができる。無向グラフでは各変数が頂点を表し、関連構モデルにおける 2 変数間の偏相間係数がゼロでない場合に頂点間を直線で結ぶ（2 変数間に関連があるとみなす)。選ばれた関連構造モデルのパラメターをチェックする場合は、メニューバーの[Fit]をクリックし[Show estimate]から選べばよい。



変数のうちで、PC,PS,DWS には

PC(地上部のリン含有量)=PS(地上部のリン含有率)×DWS(地上部の乾燥重)

の関係がある。グラフィカルモデルではこの 3 変数は互いの関連が検出されているが（下表の上三角が変数間の偏相間係数、下線付きで表す)、単純な相関係数（下表の下三角行列）では PS と PC 間の関連は検出できない。

	PS	PC	DWS
PS	1	0.97	−0.979
PC	0.135	1	0.984
DWS	−0.484	0.793	1

偏相間係数とグラフィカルモデルの応用は、変数間の見落としがちな関連を見出せる強力なツールである。

参考文献：

宮川雅巳「グラフィカルモデリング」(1997;朝倉書店)

Edward, D “Introduction to Graphical Modelling” (2000;Springer)