重回帰分析における抑制変数と多重共線性
— 相関と予測力 —

Suppressor Variable and Multi-collinearity in Multiple Linear Regression Analysis
— Correlation and Explanation Power —

榛葉 豊*
Yutaka SHINBA

Abstract : Multiple linear regression analysis is well known and easy to conceive its image of work. However, when we inquire the interpretation of explanation and prediction power of explanation variables, paradoxical aspect happens. We consider such examples, namely, occurring suppressor variable and multicollinearity. Multidimensional vector representation for sample state is discussed.

1．はじめに

来年度から数学教職科目として「多変量解析」を開講することになった．講義の構成を考える中で，数理統計学よりの，公式などの導出を含んだ構成にするか，統計ソフトを使い，実際の分析を主としたやり方か，分析の意味とか，分析結果の読み方，判断の仕方に重点を置くかなどと思いをめぐらせていた．色々の教科書を調べてみると，入門的な物は，ソフトの使い方主体のものが多く，統計的な物事の見方，推論の仕方，誤りに陥りやすい点などを分かりやすく書いてあるものはほとんど無かった．そこで，実際の分析の時，その解釈を誤りやすい事項を検討していた．

重回帰分析は，多変量解析の代表的な手法であり，そしていろいろな手法の中では意味が比較的わかりやすく，実際にもよく利用されている分析法である．しかし最近では，種々の統計パッケージ，たとえばSPSSやS-PLUSなどがPCでも手軽に利用できるようになってきて，その数学的意味の理解は勿論のこと，何をやっているのかの意味すらわからなくても，データを入力すると何らかの「分析結果」が出てしまう．簡単な回帰分析ならエクセルですら出来るのである．

目的変数をひとつの説明変数から説明，予測する単回帰分析であれば，解釈の困難はあまりないのであるが，２つ以上の説明変数を取り扱う重回帰分析では，非常に奇妙な事態や，解釈不能なことが起こる．それは「因果と相関」などという，帰納法の哲学の深刻な問題ではなく，もっと身近な，説明変数の間の相関が問題を引き起こすと言うような事である．それは主に数値的分析の不安定を引き起こすということであるが，意味合いの上での不思議さも引き起こす．

これらの困難は，単回帰分析と違って重回帰分析では相関の種類が何種類にもなることが引き起こしている．それに関する学習上の困難は，微分学の学習において，２変数以上の微分学で，偏微分が導入されるということを上回っていると思う．微分学でも全微分であるとか，斜交座標系での偏微分係数であるとか，ベクトルの共変成分，反変成分であるとか難しい．この事情と似てはいるが，重回帰分析では単なる多変数関数を扱うのではなく，理論で言えば多説明変数による目的変数の確率密度分布を，そして現実の分析ではその離散標本分布を扱うのである．多変数関数の場合の，多次元空間内の曲面ではなく，（重回帰分析では平面ではあるが）それからばらついている離散データのあらわす散布図を思い描かなくてはならないからである．

本稿では，重回帰分析の教科書に注意点としてほとんど説明なく対処法のみ書かれている場合が多い多重共線性と，あまり触れられることはない抑制変数について，その解釈を論ずる．

2010年３月11日受理
* 総合情報学部 人間情報デザイン学科

意思決定理論や確率の認知心理学の領域では，相関と説明力の関係についての「主観的定理」として取り上げられることがある．その例題として，認知心理学者，市川伸一の挙げる例[1]等を考察してみる．

2．重回帰分析における直感に反する状況

2.1 抑制変数　：　目的変数との単相関係数が正であるのに，偏回帰係数が負であるという説明変数

仮想的な例[1]を（改変してある）見てみよう．ある理科系大学で，数学と国語の入学試験の成績，それに対象受験生達の，入学後の成績を調べた．目的は，次年度の入学試験で，入学後の成績がよいであろうと予測される受験生を選別するには，どの科目の点数を重視したらよいかを考察するためであるという．

問 1「入試の数学得点と入学後の総合成績の相関は正，入試の国語得点と入学後の総合成績の相関も正である．しかし，重回帰分析を行ったら，国語の成績の偏回帰係数が負になってしまった．数学の得点が同じ受験生がいた場合，国語の点が悪い方を入学させた方がよいのであろうか」

相関が正であるという事は，数学が出来た方が入学後の成績が良い傾向があり，また国語が良い方が入学後も良い傾向があるという事になる．しかし，偏回帰係数が負と言う事は，重回帰分析で得られた重回帰方程式，すなわち目的変数を予測する超平面を規定する式である 1 次式で，その説明変数の係数が負であるという事である．これは偏微分係数と同じである．偏微分係数のときよりも，線形の式そのものに限っているのであるから，むしろ簡単ともいえる．

説明変数のうち他の変数の値が同じ個体同志なら，その説明変数の値が小さい方が良いと言う事になる．

重回帰方程式は，たとえばこの例では

入学後成績　＝　7.03×数学の点　—　6.97×国語の点　　(1.1)

である．国語の偏回帰係数は，　— 6.79　である．しかし，入学後生成と国語の成績の相関係数は　正　であるというのである．

このような事態を，国語は「抑制変数になっている」という．しかしこれは　単に　重回帰方程式において国語の偏回帰係数が負であれば抑制変数であるというのではない．重回帰方程式の係数が負で，しかも目的変数との相関係数が　正　である場合にのみ「抑制変数」というのである．重回帰方程式の係数だけなら，変数の値の正負を付け替えれば良いだけの事である．

さて，この事態は数学的には何も不思議な事はない．しかし，例の状況での感覚的不思議さと，論理的不思議さがある．

数学の点が同じ受験生が 2 人いたとき，国語が出来ない受験生を入学させて，国語の出来る学生は落第させよ，と言う納得できなさである．

もうひとつは，国語が出来ない方を優先せよということと，国語が出来た方が単相関からは優先されるという事の関係である．

前者には，以下のような説明がされることが多い．「数学の成績には，国語の能力も含まれている．従って数学の点が同じなら，国語力に助けられないでその数学の点を達成した者の方が，数学の独自能力は高い．理科系大学にはその方が向いている」．数学を生物に，国語を化学と置き換えてみれば，感覚的にもっと納得できるであろう．

2.2　偏相関係数と直接的相関，間接的相関

後者の疑問は次のように説明できるであろう．それは，偏相関係数という概念を用いればよい．相関係数は

$$r_{xy}=\frac{\sum_{i=}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}=\frac{\sigma_{xy}}{\sigma_x\sigma_y} \qquad (2.\ 1)$$

である．ここにバーが付いているのは平均値である．N はサンプル数である．

ここで，変数 X,Y に対し相関を持っている変数 Z が有るとする．その変数の分布を通して，X と Y の間に相関が出てきてしまうとする．そうすると，純粋に X と Y の間の相関関係で，Z の分布の影響を取り除いたのは，偏相関係数で，

$$r_{x,y;z}=\frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{1-r_{xz}{}^2}\sqrt{1-r_{yz}{}^2}} \qquad (2.\ 2)$$

で，定義される．（この式は次項で述べるベクトル図による表現で，ベクトル $\vec{x}$ と $\vec{y}$ の角度のコサインが相関係数であるのに対し，$\vec{z}$ の影響を取り除くため，$\vec{z}$ 方向成分を引き去ったもの同志の角度のコサインである）．

抑制変数という事態が起こっている場合でも，すなわち，国語と入学後の成績の相関が正であってそれなのに偏回帰係数が負である場合でも，この偏相関係数は X,Y,Z をそれぞれ，国語，入学後，数学として，負になりうる．偏相関係数が負で，偏回帰係数も負と言うことになる．これなら納得できるであろう．

本来，2 つの変数が持つ相関と，仮に 2 つの変数の間に

本質的な相関関係が無くても，別の事情により相関を持ってしまった場合である．後者は，メカニズム的には関係ない変数なのに，偶然相関があるようなサンプルになってしまったというのではなく，その２つの変数とは別の第３の変数との相関を通して関係が出来てしまったという場合である．

直接的相関と間接的相関を考えてみる．変数間の相関のうち，１つの説明変数だけを変化させて，他の説明変数を固定した場合の相関が直接的な相関である．一方１つの説明変数を変化させた場合，別の説明変数の値も変化させた説明変数と相関していれば変わってしまう．この別の説明変数と目的変数が相関していれば，それを通しての目的変数の変化が起こる．これが間接的な相関である．偏相関係数は，前者のことである．一方普通の相関係数は，直接的な相関と間接的な相関を足し併せたものである．

問１の状況を考えてみるために，３次元の散布図を思い描いてみよう．市川は個体データの例とその散布図を全く示さない[1]．そしてすぐに，次項で述べるベクトル図に逃げてしまう．この問題の散布図を想像する事は３次元の問題であって，その上どの角度で見ても見にくい点の分布になり，少し難しい．しかし，筆者が知るかぎり示されるのを見た事がない３次元散布図は，この問題の理解のために，ベクトル図と同等以上に重要であると考える．それは以下のようになるであろう．

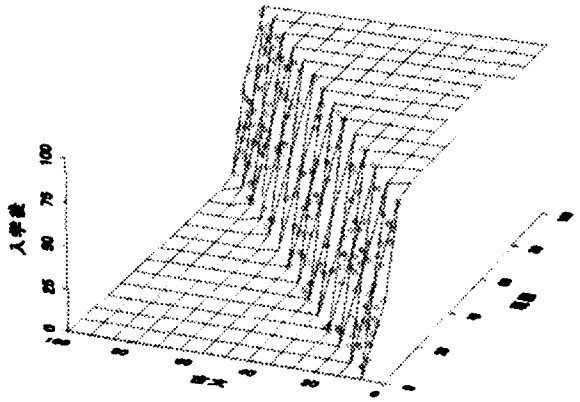

Fig-1　３次元散布図

この問１の状況は，数学と国語を水平面の直交座標で表し，縦軸で入学後の成績を表した散布図で表される．数学と国語は正の相関であるから，Fig-1 の手前右から左上方向の直線に沿って散布点の射影がばらついている．その上方に個体の散布点があるわけである．重回帰平面は直前に述べた斜めの線から数学の軸の方に倒れるような形で存在する．図の崖状の面である．この面の傾きだと，数学の偏回帰係数は正，国語の偏回帰係数は負になる．

それでは，この３次元散布図を念頭に入学後と国語の間の相関係数と偏相関係数の符号が違いうるという事を理解してみよう．偏相関係数は，第３の変数（数学）との相関がなければ単相関係数に一致する．これは重回帰平面近傍に一様に点がばらまかれているような場合である．第３の変数との相関があるときには，重回帰平面上の水平面で見て放射方向に散布点が分布している．それは数学―入学後平面への射影で見ても，国語―入学後平面への射影で見ても同様である．

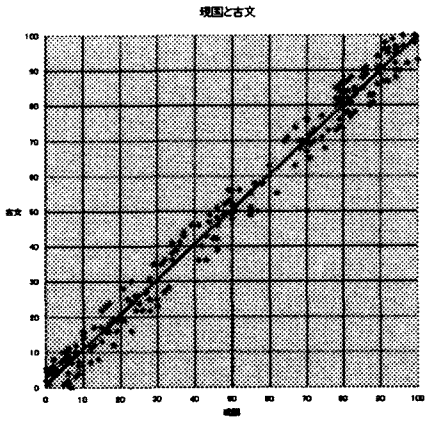

Fig-2　説明変数平面への射影

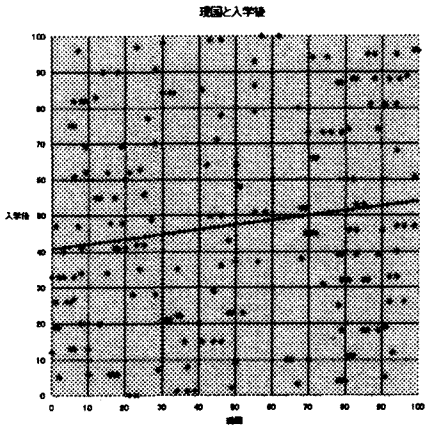

Fig-3 国語―入学後平面への射影

ここで，国語―入学後平面への射影，すなわち２次元散布図で見れば相関が正なのに，なぜ偏相関係数が負に成るのかという事が問題である．これは一言で言えば数学の点数をある一定値に固定した数学の軸に直交する Fig-3 の面に平行な平面で考えるという事を，その固定する値を色々変えた平面群で考えれば理解できる．数学の値を固定した垂直な平面と重回帰平面の交線は，国語の点が増すと入学後が下がるという傾きになっている．これが，偏相関係数が負だという意味である．（Fig-3 に書かれている右上がりの直線は，単回帰直線である．）この事は，数学の固定す

る値が何点であろうと同じ事情である.

言葉を改めて述べよう. 数学の点を特定しないで全体としてみれば, 国語の点が高い方が入学後の点も高い. Fig-3の単回帰直線である.

しかし, 個別に数学の点が同じ2人でという比較になると, 統計学の言葉で言えば層別の比較という事になる. そこでは国語の点が低い方が入学後の成績は高いという事になるのである.

国語 — 入学後の2次元散布図で, 数学の得点の層別に散布点を色分けしてみればすぐ納得がいくであろう.

違う例で述べてみよう. コーラAとBの都市ごとの売り上げを考えよう. ある都市でのAの売り上げとBの売り上げを1個体のデータとする. すると, 全国の都市の統計では右上がりの正の相関になったという. しかし, 人口別に層化して, 例えば中規模都市のなかでの統計では, AとBは相関係数が負で反相関していたという. これは都市の規模で, コーラ全体の売り上げは大体決まっていて, A+Bはほぼ一定であったためである. この層別相関と全体での相関の関係と似た事情である.

2.3 多次元ベクトル図

市川は, このような事態をはじめ, 変数間関係を考えるのに, 多次元ベクトル図の使用を薦めている[1].

例えば2変数の重回帰分析で言えば, 散布図は3次元である. そこに, サンプル数N個の散布点がばらまかれているわけである. これを, 統計力学の位相空間と同じ考えで表現してみる. 3次元空間の散布の状態は, 3個の変数の値をN個並べたもので表される. すなわち

$$\begin{matrix}(x_1, y_1, z_1)\\ \cdots \\ (x_N, y_N, z_N)\end{matrix} \qquad (2.3)$$

である. これを縦方向に読んでN次元空間の3つのベクトル

$$\begin{aligned}\vec{x} &= (x_1, x_2, \cdots x_N)\\ \vec{y} &= (y_1, y_2 \cdots y_N)\\ \vec{z} &= (z_1, z_2, \cdots z_N)\end{aligned} \qquad (2.4)$$

で表す. 特に生データではなくて, その変数の平均値からの偏差

$$\vec{x}' = \left((x_1 - \bar{x}), (x_2 - \bar{x}), \cdots (x_N - \bar{x})\right) \qquad (2.5)$$

を用いるとなお便利になる (以下プライムは省略する). こうすると, 相関係数 (2.1) はN次元空間の内積で表される.

$$r_{xy} = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|} \qquad (2.6)$$

すなわち, N次元空間での$\vec{x}$と$\vec{y}$の「角度」θのコサインが相関係数である事になる.

$$\vec{x} \cdot \vec{y} = \|\vec{x}\| \|\vec{y}\| \cos\theta \qquad (2.7)$$

ベクトル図では変数の状態がベクトルで表されその角度が相関係数になっている. またベクトルのノルムを$\sqrt{N}$で割ったものはその変数の標準偏差になっている.

$$\sigma_x = \frac{1}{\sqrt{N}} \|\vec{x}\| \qquad (2.8)$$

重回帰分析とは目的変数$\vec{z}$を説明変数$\vec{x}$と$\vec{y}$の線形結合で予測しようとする事である. 言い換えると, $\vec{x}$と$\vec{y}$で張られる面への, $\vec{z}$の射影を求める事である.

多次元ベクトル図で抑制変数を考えてみると次のようになる.

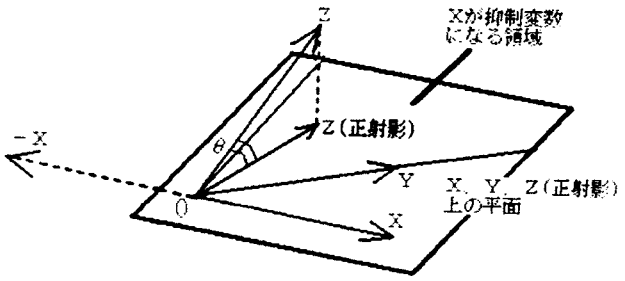

Fig-4 多次元ベクトル図

数学Yと国語Xのベクトルが張る平面上で数学と国語のベクトルは90度以内の角で交わっている (相関が正である). 一方目的変数である入学後のベクトルZは, 一般にはその平面から浮いているが, その先端のこの平面への正射影が重回帰方程式に対応する. Xが抑制変数という事態では, 国語Xの偏回帰係数が負なのであるから, この射影が, 数学と国語がつくる狭角の中に落ちないで, 外に落ちるという事である. しかし, 国語の軸の負の方向側には落ちない. なぜなら, 国語の軸の負の方向の反平面になると, 国語と入学後のベクトルの角度が90度以上と成って, 相関係数が負になってしまうからである.

つまり, 先の数学と国語がつくる角の外で, 国語の軸と直交する直線との間の領域でのみ, 抑制変数という事態が

起こるのである.

重回帰分析を行う際の注意点として,いくつもの教科書に書かれているのは,(後述の)多重共線性(複数の説明変数間に高い相関がある事)という事態であり,それはこの抑制変数と一体のように書かれている.しかしこの2つは独立な現象である.この点を間違えている教科書がある.ただ,多重共線性が起こっている場合には,この数学と国語の角が非常に小さくなるという事だから,抑制変数が起こりやすいとはいえる.

2.4 目的変数と相関がない変数は予測の役に立たないか

次の問は,普通の多変量解析の教科書でも比較的よく取り上げられている.

問2「目的変数との単相関がゼロの説明変数は,予測の役に立たないと言って良いか」

数学と入学後の相関はゼロ,物理と入学後の相関は正であるとしよう.説明変数間の内部相関があれば,目的変数との直接の相関がゼロであっても,目的変数と相関を持つ他の説明変数を通して,影響力を持てるのである.2.2項で述べた間接的な相関である.偏相関係数は有限の値を持ちうるし,他の説明変数と組み合わせれば,その他の変数だけの場合の予測力を向上させる事が出来る.数学と入学後は相関していなくて,数学 — 入学後 平面への射影が傾向を持たなくても,数学と物理に正の相関があれば数学が出来ると物理が出来る傾向があり,従って入学後の成績はよいであろうとなる.

ベクトル図で考えてみよう.入学後のベクトルは数学のベクトルと直交している.入学後のベクトルと 数学 — 国語 平面との角度は,入学後のベクトルと国語のベクトルとのなす角より,一般には小さくなりうる.と言う事は国語だけを説明変数にした単回帰分析での予測より,数学を加えた重回帰分析の方が,相関係数が高くなりうるのである.

2.5 多重共線性

前項の拡張になっているが,重相関係数が非常に大きい場合を考えてみる.

問3「入試の数学と入学後の成績の単相関係数 0.14
入試の国語と入学後の成績の単相関係数 0.0
入試の数学と国語の単相関係数 0.99
であったという.
このとき重相関係数はどのぐらいと思うか」

問題文を,直感に従って考えてみる.説明変数と目的変数の間の相関係数からは,「数学は余り役に立たない変数」,「国語は全く役に立たない変数」である.しかも数学と国語は「ほとんど同じ」変数であり,片方で良いと思われる.従って,ほとんど同じ役に立たない変数からは,ほとんど目的変数の予測は出来ない.重相関係数はゼロに近いと推測出来るだろう.

ところが驚くべき事に,重回帰分析すると,重相関係数は 0.992 すなわちほとんど完全に予測できる,と言う事が起こっているというのである.

市川はここでも個体データと散布図を示さない.しかしこの状況をFig-1の3次元散布図で考えてみよう.それは,2.2項で考えた散布状況で,重相関係数がほとんど1であるから,散布点が重回帰平面にへばり付いている様子を想像すればよい.これは2変数関数(それも線形の関数の)微分学とほとんど同じである.その上で問3の相関係数を考えてみると次のようになる.

入学後—数学平面への3次元散布図からの散布点の正射影には傾向がない.同様に入学後—国語平面への正射影も全く傾向がない(Fig-3).一方,実はこの問題での重回帰方程式は,式1.1なのである.すると Fig-2 の斜めの線の上側から立ち上がった重回帰平面にへばり付いた散布点は,垂直に近いのであるから数学—国語平面上への正射影の分布で考えて,放射方向の直線からほとんど外れていない(Fig-2).なぜなら,数学と国語の単相関は 0.99 だからである.

このような状況では,データのわずかな変化で,重回帰平面が敏感に大きく変動したりする.この現象を 多重共線性という.説明変数間に高い相関が見られる場合である.多重共線性はどの本にも書いてあって,その様な変数を用いた重回帰分析はそのままでしてはいけないと書かれている.

一般に多重共線性の発生している状況では,重回帰平面の傾きが極端に大きかったり,重相関係数がほとんど1になったりする.そして,重回帰平面は非常に不安定である.

この多重共線性の状態を多次元ベクトル図で言えば,数学と国語の方向がほとんど同じな場合である.問3では重相関係数も1に近いので,その状況に更に加えて,入学後ベクトルも,数学—国語平面に低く張り付いている.(数学,国語の両ベクトルとの角度は必ずしも小さくはない).

今の状況を考えると,数学の点と国語の点の差で入学後の点が決まっているという状況である.従って,数学と国語という2つの変数で考えるより,

数学 — 国語

という新しい説明変数を考えた方がよい.主成分分析の第1主成分と同じ事である.このとき第2主成分に相当するのは

数学 + 国語

である．しかしそうしても，非常に不安定なことは変わりがないから，それ以上の分析はやはり避けた方が良い．現実には問3のような状況は起こりにくいとは思われる．

3．まとめ

重回帰分析は，説明変数の持つ説明力を吟味したりする幾つかの多変量解析の手法のうちでも比較的分かりやすい手法であろう．しかし，それでも直感に反するような事例が起こる．相関係数が持つ意味合いや，多変量である事から来る，影響力のネットワークなどの分かりにくさもある．

特に相関という事が，直接の変数本来の相関の他に，第3の変数を通しての間接的な相関という事態もあるということが分かりにくいだろう．これは微分学でも見られた事である．可微分決定論的モデルで $z = f(x,y)$ において $y = g(x)$ であるとする．これは，相関係数の値が1に近いときの，2つの説明変数間の内部相関を近似的に表している．すると，

$$\frac{dz}{dx} = \frac{\partial z}{\partial x} + \frac{\partial z}{\partial y}\frac{dy}{dx} \qquad (3.1)$$

である．たとえ，$\partial z / \partial x$ がゼロ，すなわち目的変数と説明変数 x の相関がゼロにあたる，としても，第2項で表される説明変数間の内部相関を通しての相関が有り得るのである．この事の理解は逆に多変数の微分学の理解にも役立つのではないであろうか．

また，多次元ベクトル図による理解も，数学的には自明な事であるのに，なぜか相関を扱う局面での，多変量解析の入門的教科書では説明にあまり用いられていない．多次元ベクトル図は積極的に援用されるべきであろう．

但し多次元ベクトル図は，変数自体の性質を表現しているのではなく，「変数の値がとるサンプル」を表すベクトルの関係図である．すなわちそこに書き込まれているのは，点数空間の状態ベクトルである．これを忘れると理解が難しくなる事に注意しなくてはならないだろう．

因果と相関という難問はさておいても，相関という事はどういう事かよく理解しなければならない．また回帰係数と相関係数の変数の持つ説明力という事に関する意味の違い，回帰の片方向性（相関は2変数間の対称な関係である）ということも忘れてはならない．

謝辞

本研究に関連した図を作成してくれた，静岡理工科大学の小林真一氏に感謝します．

参考文献

1） 市川伸一，「決定における規範的理論と直感的推論」，小橋康章 著『決定を支援する』，東京大学出版会（1988年）

2） 柳井晴夫，岩坪秀一，『複雑さに挑む科学』，講談社（1976年）

3） 君山由良，『重回帰分析の利用法』，データ分析研究所（2004年）