多変量解析を理解するために(2)

重回帰分析の基礎

高木 廣文 | Takagi Hirofumi

東邦大学医学部看護学科国際保健看護学研究室教授

■1950年生まれ。74年東京大学医学部保健学科卒。79年同大大学院博士課程修了。同年米国国立衛生院奨励研究員として米国国立環境保健学研究所勤務。81年聖路加看護大学講師、82年同大助教授。89年文部省統計数理研究所助教授。99年新潟大学医療技術短期大学部看護学科教授、同年同大医学部保健学科教授。2006年4月より現職。著書に『ナースのための統計学 — データのとり方生かし方』(医学書院)、『多変量解析ハンドブック』(現代数学社)、『健康科学とコンピュータ』(共立出版)など。

1. 相関関係と回帰分析

たばこの吸いすぎは肺癌のもと、塩分の摂りすぎは高血圧のもと、肉・油脂などの摂りすぎは脳卒中の原因の1つ、のように、我々の生活する世界の中の物事には相互に何らかの関係があるのが普通である。

このような関係に科学的に注目したのは、ゴルトン(Golton, F., 1822~1911)である。ゴルトンは、スイートピーの親種と子種の観察を行い、大きな親種の子種の平均値は、親種の平均値より小さくなり、小さな親種の子種の平均値は親種の平均値より大きくなるような傾向があることを見つけた。この傾向は「reversion(先祖返り)」と初めは呼ばれていた。さらに、ゴルトンは、両親の身長と子供の身長の関係を調べた結果、大きい両親から生まれた子供の身長は大きく、小さな両親からの子供の身長は小さい傾向があることを見つけた。彼は、このような関係を「**相関** correlation」と呼んだ。しかし、身長の高い両親の子供がさらに大きくなるわけではないし、逆に小さい両親の子供がさらに小さくなるというわけでもない。全体として、平均値に向かう傾向があることを見つけ、この平均化の傾向を示すために「**回帰** regression」という言葉が初めて用いられた。

現在では、ある変数について、他の幾つかの変数を用いて予測などを行うための統計的な手法は「**回帰分析** regression analysis」と呼ばれている。

前回説明したように、回帰分析では、予測すべき変数は「基準変数」、予測のために使用する変数は「説明変数」と呼ばれる。

現在使用されている「(単)相関係数」は、ゴルトンの弟子のカール・ピアソンによって提唱されたもので、「ピアソンの積率相関係数」と正式には呼ばれている。ほとんどの多変量解析の手法が、この2変数間の相関関係を示す単相関係数を基にしている。回帰分析も当然、この単相関係数を基にして解析を行うのが普通である。回帰分析で、説明変数が1つの場合には、「**単回帰分析**」と呼ばれる。説明変数が2つ以上ある場合には、「**重回帰分析**」と呼ばれる。今回は、重回帰分析の基礎的な考え方について解説しよう。

2. 重回帰分析の目的と線形モデル

重回帰分析を行う目的は、大きく分けて2つある。すなわち、

①幾つかの説明変数を用いて、ある基準変数に関する予測式を求める。

②幾つかの説明変数のある基準変数に対する影響の程度を検討する。

ことである。

父親と母親の身長から、子供の身長を予測する式を求める場合が①に該当する。また、年齢、1日当たりの食塩摂取量や脂肪摂取量などが、血圧にどのような影響を及ぼすのかを検討するのならば、その分析は②が目的ということになる。

実際には、血圧は簡単に測定できるので、わざわざ調査してデータを集め、予測式を求める必要はないだろう。現実には、重回帰分析により予測式を求めても、その予測式自体には価値がない場合がほとんどである。重回帰分析を用いるのは、ある変数への他の変数の影響を検討するためというのが普通であろう。

それでは、重回帰分析の理論の基礎的な部分をごく簡単に説明しよう。今2つの変数をXとYとし、変数Xの値xから、変数Yの値yを予測するための式を考えよう。最も簡単な式は、説明変数の値xに適当な数値bを掛け、これに一定の数aを加えるという方法である。

すなわち、基準変数Yの予測値を$\hat{y}$としたとき、

予測値$\hat{y}$＝[係数bの値]×[説明変数の値x]＋[定数a]

によって予測を行うことにする。これを記号のみで書けば、

$$\hat{y} = bx + a$$

となる。

上式は、「**回帰式** regression equation」と呼ばれる。また係数bは「**回帰係数** regression coefficient」と呼ばれている。縦軸をYとし横軸をXとした場合に、この数式が表している1次直線は、「**回帰直線** regression line」と呼ばれている。回帰分析で求めるべきものは、係数bの値と定数aの値である。

一般に、説明変数がp個ある場合、すなわち、説明変数がX_1, X_2 ..., X_pとあるとする。そして、説明変数の観察されたデータが、x_1, x_2 ..., x_pであるとする。このとき、基準変数の予測値$\hat{y}$を、

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

のように、それぞれの説明変数に重みを掛けて、その和を計算するという式を考える。ここで、b_0は定数項であり、他の変数への重みの記号表記に合わせたものである。

このような重回帰分析のためのモデル式は、一般には「**線形モデル** linear model」と呼ばれている。重回帰分析では、この基本式は「**重回帰式** multiple regression equation」と呼ばれている。

重回帰分析では、上式の重みを求めることが解析の主目的になる。各変数に掛ける重みは「**偏回帰係数** partial regression coefficient」と呼ばれている。

重回帰分析の目的②である、説明変数の基準変数への影響の有無を検討するには、それぞれの説明変数の偏回帰係数が0か否かを検定することで調べることができる。

また求めた重回帰式の予測の、基準変数の予測値と実際値との相関係数である「**重相関係数** multiple correlation coefficient」の大きさで判断できる。なお、重相関係数の2乗は「**多重決定係数** multiple coefficient of determination」と呼ばれ、分析に用いた説明変数により、どの程度の予測が可能かを示す指標になる。

3. 臨床検査データの解析例

まず、重回帰分析の実際例を示しながら、重回帰分析の理論的な背景も簡単に説明しよう。ここでは、約1,200名の生理学的な臨床検査データ[1]を基に、簡単な解説をしよう。

表1は、分析に使用する6項目、すなわち年齢、性、BMI、身長、体重、および皮下脂肪厚の平均値と標準偏差を示したものである。すべてのデータを完備していたのは1,066例であった。多変量解析を行う場合には、分析に必要なデータの欠損しているケースが実際にはかなりある。ここでは、すべての変数のデータが揃った例のみを扱うものとする。

なお表中のBMIとは、Body Mass Indexの略号であり、最近では肥満度の指標としてよく用いられている。BMIの値は、体重(Kg)／身長(m^2)により求められ、BMIが22の場合を標準とし、それ以上ならば肥満傾向を、それ以下ならば痩せ傾向を示すことになる。また、性は男なら1、女なら2を数値データとして与えている。一般には、性などのような質的変数を多変量解析に含める場合、ダミー変数（適当な個数のカテゴリに0, 1を与えて分類する）を用いて分析を行うのが普通である。しかし、カテゴリ数が男女のように2つの場合には、1, 2のような数値のままで分析に用いても問題はない。

ここで基準変数は皮下脂肪厚とし、残りの5つの変数を説明変数とする。

表2は重回帰分析に用いる6変数間の単相関係数（ピアソンの積率相関係数）を示したものである。無相関の検定の結果は示していないが、1,066例の場合には、相関係数が0.06を越えると5％水準で有意となる（つまり、母集団での相関係数は0ではないと言える）。従って、表2を見るとBMIと性、身長以外はすべて有意な相関があることになる。表2の単相関係数を見ると、皮下脂肪厚と最も相関が高いのはBMIであり、次に性、体重の順になっている。また、年齢と身長は皮下脂肪厚とは逆相関になっている。さらに説明変数間の相関係数をよく見ると、BMIと体重は0.803という極めて大きな数値になっている。これは、BMIの計算方法から言って、至極当たり前のように思えるかも知れない。しかし、身長とは−0.018とほぼ無相関となっており、必ずしもBMIの計算式から予測されるような大きな逆相関を示しているわけではない。

表2の結果も詳細に検討すると、上記のように興味深い点を見つけることができるが、個数が多くて煩雑である。皮下脂肪厚と他の説明変数間の関係を検討するために、重回帰分析を行ってみよう。

表1　分析に用いた変数の平均値と標準偏差(1,066例)

変数名	平均値	標準偏差
1) 年齢（歳）	55.6	9.2
2) 性	1.3	0.4
3) BMI（Kg／m^2）	24.8	3.4
4) 身長（cm）	164.0	8.1
5) 体重（Kg）	66.7	11.2
6) 皮下脂肪厚（mm）	15.4	7.7

表2　変数間の相関係数行列

変数名	1)	2)	3)	4)	5)	6)
1) 年齢	1.000	0.112	−0.141	−0.344	−0.327	−0.203
2) 性	0.112	1.000	0.059	−0.674	−0.347	0.472
3) BMI	−0.141	0.059	1.000	−0.018	0.803	0.603
4) 身長	−0.344	−0.674	−0.018	1.000	0.571	−0.245
5) 体重	−0.327	−0.347	0.803	0.571	1.000	0.351
6) 皮下脂肪厚	−0.203	0.472	0.603	−0.245	0.351	1.000

表3は、皮下脂肪厚を基準変数に、他の5変数を説明変数として重回帰分析を行った結果を示したものである。

表3の「偏回帰係数」とは、前述したように各説明変数に掛ける重みである。括弧に入っている「標準誤差」とは、偏回帰係数の推定上の精度を示すもので、後述するように検定のときに使用する。

「標準偏回帰係数」は、すべての変数が平均0、分散1に標準化した場合の偏回帰係数を示すものである。「偏相関係数」は、皮下脂肪厚と各説明変数の関係について、他の変数の影響を除いた場合の相関係数を示すものである。

基準変数に対する各説明変数の影響を簡単に見る場合には、標準偏回帰係数か偏相関係数を見ればよい。絶対値が大きいほどその変数の基準変数への影響が大きいものと考えられる。ただし、表3でも分かるように、両者の値は必ずしも一致するわけではない。このため、母集団での偏回帰係数（母偏回帰係数）が0か否かの検定が必要であるが、この点については、次回に説明する。

基準変数の予測値は、偏回帰係数と説明変数の積の合計に、定数を加えた値によって求められる。重回帰式による予測値と実際値との相関係数は「重相関係数 multiple correlation coefficient」と呼ばれており、この値が1に近いほど予測がうまく行くことを示す。ところが、重相関係数は説明変数の増加に伴い単調に増加し、説明変数の個数が「ケース数−1」になると必ず1となる。従って、重相関係数はケース数が説明変数の個数に比べて大きくない場合、あまり適切な指標とは言えなくなる。そのような場合、ケース数と説明変数の個数を用いて、重相関係数を修正した「自由度調整済重相関係数」を求めればよい。実際にはさらに、自由度の再修正を行った「自由度再調整済重相関係数」を求めることが多い。これらの2つの重相関係数は、説明変数の個数がある一定の値を超えると減少し、自由度再調整済重相関係数は負の値さえ取り得る。もしも、その値が負となれば、求めた重回帰式には何の意味もないことになる。

表3 皮下脂肪厚の重回帰分析の結果

変数	偏回帰係数（標準誤差）	標準偏回帰係数	F値（p値）	偏相関係数	ΔRSq
年齢	−0.139 (0.018)	−0.166	58.2 (0.000)	−0.228	0.023
性	8.450 (0.480)	0.479	310.1 (0.000)	0.476	0.121
BMI	0.850 (0.336)	0.370	6.4 (0.011)	0.078	0.003
身長	−0.096 (0.103)	−0.100	0.9 (0.353)	−0.029	0.000
体重	0.154 (0.123)	0.223	1.6 (0.211)	0.038	0.001
定数	−3.196				
重相関係数（2乗）[注]		0.765 (0.586)			
自由度調整済重相関係数		0.764 (0.584)			
自由度再調整済重相関係数		0.763 (0.582)			

注）F(自由度) = 299.6(5, 1060)，$p < 0.0001$。

表3の最右欄の「ΔRSq」は、その説明変数を重回帰式から除いた場合、どれだけ重相関係数の2乗が減少するかを示すものである。従って、この値が大きいほど基準変数に与える影響が大きいことを示す。実際には、この値を用いて検定のためのF値が求められている。なお、ΔRSqは重回帰式からある変数を1つ除いた場合の多重決定係数の減少分であるので、すべての変数についてΔRSqを合計しても、多重決定係数と一致しないので注意がいる。

次回は、重回帰分析での検定と説明変数の選択の問題について解説する。

＊参考文献
[1] 高木廣文(2001)：生理学的データの重回帰分析：超音波検査技術, 26(1), pp.28-34.
[2] 柳井晴夫・高木廣文編著(1986)：多変量解析ハンドブック：現代数学社, 京都.
[3] 高木廣文・柳井晴夫編著(1995)：HALBAUによる多変量解析の実践：現代数学社, 京都.
[4] 高木廣文(2006)：HALBAU7によるデータ解析：シミック㈱.