多変量解析を理解するために(3)

重回帰分析での変数選択

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. 重回帰分析における説明変数の選択の必要性

前回説明したように、重相関係数は説明変数の個数が増加するにつれ、その値も単調に増加し、説明変数の個数が標本数より1少ない数に達すると、その値は必ず1になる。

従って、説明変数の個数を増やすことで、重回帰分析による予測の程度を見掛け上、幾らでも良くすることができる。このため、単なる見掛けだけの良さを避け、分析に最適な説明変数のみを重回帰式に含めるための適当な方法が必要となる。今回は、重回帰分析での説明変数の選択方法と、そのための基準について簡単に説明する。

2. 総当たり法

総当たり法とは、重回帰分析で説明変数が複数個ある場合、可能な説明変数の組み合わせすべてについて重回帰式を求め、それらを比較し、最良の変数の組み合わせを選ぶ方法である。例えば、基準変数Yに対して、説明変数がA、B、Cと3つある場合を考えてみよう。説明変数を1つ使う場合、A、B、Cのそれぞれについて、3とおりの単回帰式が求められる。説明変数を2つ使う場合、AとB、BとC、CとA、の3とおりの重回帰式が求められる。そして、A、B、Cの3つすべてを使う重回帰式が1つあり、全部で7とおりの重回帰式を比較すればよい。どの重回帰式が優れているかという選択の基準は、自由度調整済重相関係数か自由度再調整済重相関係数（こちらが普通）が最大となる組み合わせを採用すればよい。

しかし、単純な計算で分かるように、説明変数の組み合わせの個数は、説明変数がk個あれば、(2^k-1)とおりある。例えば、説明変数が10個ある場合は、説明変数が1つの単回帰式から、10個すべてを用いる重回帰式まで、全部の説明変数の組み合わせ個数は1,023とおりある。このように、説明変数の個数の増加とともに、比較すべき説明変数の組み合わせによる重回帰式の個数は急激に増加する。時間がある場合には、パソコンをフル稼働させて計算させてみるのもおもしろいかも

知れないが、実際の解析では、総当たり法は、説明変数の個数が少ない場合のみに用いられる。

3. 逐次選択法

説明変数の選択の基本的な方法は、大きく分けると2とおりになる。すなわち、適当な基準により説明変数を1つずつ重回帰式に加えていく「**変数増加法**」と、すべての説明変数を用いた重回帰分析から、逆に1つずつ変数を減らしていく「**変数減少法**」である。

変数増加法は手順が簡単なのでよく用いられる方法である。しかし、重回帰式に一度含めた説明変数は、新たな説明変数が追加されることで、基準変数への説明力がほとんどなくなってしまっても、重回帰分析から除外されることがない。この点を考慮して、重回帰式に一度含めた説明変数であっても、ある基準を満たさなくなった場合に、重回帰式から除外する方法は、「**変数増減法**」と呼ばれている。

同様に、変数減少法において、重回帰式から一度除外された説明変数であっても、他の説明変数が重回帰式に追加されることで、それらの説明変数間の相関関係により、基準変数に対する寄与が高くなることもある。そのような場合に、一度除外された説明変数でも再度採用する方法は、「**変数減増法**」と呼ばれている。

これら4つの方法による説明変数選択による重回帰分析の結果は、変数相互の相関関係に大きく依存している。すべての方法が同一の結果となる場合もあるし、全く異なる場合もある。一般には、変数減増法が最も良い結果を与えると言われている。

4. 説明変数選択の基準

説明変数選択の基準には、偏回帰係数の検定のF値を用いることが多い。すなわち、用いた説明変数の基準変数に対する寄与が0か否かという検定は、「母偏回帰係数=0」という帰無仮説の検定を行うことである。

今、k個の説明変数が重回帰式に含まれており、そのときの重相関係数をR_kとする。ある説明変数を1つ取り除いた場合の重相関係数をR_{k-1}とすると、取り除いた変数の影響は、

$$F_0 = \frac{(n-k-1)\times(R_k^2 - R_{k-1}^2)}{1-R_k^2}$$

により求められるF値が自由度$(1, n-k-1)$のF分布に従うことから検定できる。

表1は、前回掲載した重回帰分析結果を再掲したものである[1][2]。表1中のF値が各説明変数の寄与の程度を表しており、( ) 内のp値が検定のための有意確率を表している。なお、上記のF値は、

$$F_0 = \left[\frac{\text{偏回帰係数}}{\text{標準誤差}}\right]^2$$

によっても、簡単に求められる（ただし、表1中の数値は丸められているので、電卓で計算すると結果は完全には一致しない）。

表1　皮下脂肪厚の重回帰分析の結果（前回表3の再掲）

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)，$p < 0.0001$。

通常の検定では、有意水準として５％が用いられることが多い。しかし、重回帰分析での変数選択の基準としてｐ値を５％に設定するとは限らない。

説明変数選択の基準としては、適当な大きさのＦ値（F_{in}）を設定し、検定のためのＦ値であるF_0と基準値F_{in}の数値を比較することで、その説明変数を重回帰式に含めるか否かを決めるようにする。すなわち、ある説明変数のＦ値が、$F_0 > F_{in}$であれば、重回帰式に含めることにすればよい。この場合、重回帰式に含まれていない説明変数が複数個ある場合には、最大のＦ値を持つ変数を１つだけ選択するようにする。もし、すべての変数のＦ値がF_{in}より小さければ、そこで説明変数の追加を終了すればよい。これが、Ｆ検定による変数増加法の手順である。

変数減少法も同様の基準を用いて行うことができる。すなわち、すべての説明変数のＦ値を求め、その最小のものをF_{min}とする。F_{min}の値があらかじめ設定しておいた変数除外のための基準値F_{out}と比べて小さければ、その説明変数を重回帰式から除くことにすればよい。すべての説明変数のF_0の値が基準値F_{out}より大きければ、そこで説明変数の減少を終了すればよい。

変数増加法や変数減少法の基準であるＦ値をどのくらいにするかは、最終的に重回帰式に残される変数の個数に大きく影響する。普通は、F_{in}とF_{out}の値には２を用いることが多い。この２つの基準の数値を異なる値にすることは滅多にない。

変数増減法では、新たに説明変数を１つ選択した後で、重回帰式の他の説明変数についてF_0を求める。その中で最小の数値であるF_{min}が、$F_{min} < F_{out}$の場合、その説明変数を重回帰式から除くことにすればよい。

同様に、変数減増法は、説明変数を１つ重回帰式から除いた後で、既に重回帰式から除かれた他の説明変数についてF_0を求め、その最大の数値であるF_{max}が、$F_{max} > F_{in}$ならば、その説明変数を再び重回帰式に含めるものとすればよい。

表１を見ると、身長のＦ値が0.9と、２より小さいので、まず重回帰分析から除外される。数値は示していないが、次に身長が除外され、追加される説明変数はなかった。その結果、表２のような結果が得られた[1]。なお、変数減増法で注意すべき点としては、表１のようにＦ値が２未満の説明変数が２つ以上ある場合でも、必ず変数の削除は１つずつ行うということである。これは、各変数選択の段階でのＦ値が、変数相互間の相関構造によって決まってくるからである。ある変数を１つ除いたり加えたりすることで、それまで有意でなかった説明変数が、突然有意になることはよくあることである。

表２では、説明変数として年齢、性、BMI（肥満度）の３変数が最終的に選択された。なお、「性」は男＝１、女＝２、として解析した。Ｆ値を見ると、表１に比べてすべての説明変数で大きくなっていることが分かる。特に、BMIは標準偏回帰係数が0.370から0.550、Ｆ値が6.4から755.0と最大になっていることが分かる。偏相関係数も0.078から0.645と大きく変化している。これは、前回示したように、BMIと体重の間には0.803という極めて大きな相関があることによるものと考えられる。

重回帰分析では、偏回帰係数の推定のために相関係数行列の逆行列を計算する必要がある。このため、説明変数の間で相互に相関の高い変数が存在すると、偏回帰係数の推定値

が不安定になり、あまり信頼の置けない結果を生じることがある。そのような場合、基準変数との相関係数の符号に関係なく、一方の説明変数の偏回帰係数が正であれば、他方の説明変数については負の値になることが多い。このような状況は、「多重共線性の問題」として知られている。この問題を避けるには、簡単な方法として一方の説明変数を除けばよい。しかし、3変数以上が相互に相関が高い場合には、問題が厄介であり、一般的な解決方法はない。主成分分析（後述）などを用いて、合成変数を求めてから、重回帰分析を行うなどの工夫をするのも解決方法の1つである。

表2　皮下脂肪厚についての変数選択後の重回帰分析結果

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	−0.148（0.017）	−0.177	77.3（0.000）	−0.260	0.030
性	8.106（0.352）	0.460	530.1（0.000）	0.577	0.207
BMI	1.265（0.046）	0.550	755.0（0.000）	0.645	0.295
定数	−17.935				
重相関係数（2乗）[注]		0.765（0.585）			
自由度調整済重相関係数		0.764（0.583）			
自由度再調整済重相関係数		0.763（0.582）			

注）F（自由度）＝498.2（3, 1062），p＜0.0001。

このようにして求められた重回帰式の妥当性は、重相関係数の検定によって確認できる。帰無仮説は「母重相関係数＝0」であるが、この仮説が棄却できないような場合は滅多にない。標本数がn、説明変数がk個、重相関係数がR_kの場合、

$$F_0 = \frac{(n-k-1)\times R_k^2}{k\times(1-R_k^2)}$$

を求める。上記のF_0が自由度（k, $n-k-1$）のF分布に従うことを用いて検定できる。

実際に表1も表2も高度に有意であることが分かるだろう。なお、表1と表2を比べると、重相関係数と自由度調整済重相関係数は若干減少したものの、自由度再調整済重相関係数は変化していないことが分かる。この結果は、求めた重回帰式がより適切なものになったことを示していると考えてよい。

表2の結果から、皮下脂肪厚は、BMIが大きいほど、男性と比べて女性ほど、年齢は若年者ほど、その値が大きいことになる。一般には、年齢が高いほど肥満傾向があるように思われるのだが、ここで用いた対象者では、高齢者ほど皮下脂肪厚が薄いようである。これらの結果は基本的には単独の変数での解析と大差はないが、複数の変数を組み合わせることで、基準変数の予測の程度はより高められている。

5. 計算には

重回帰分析は、多変量解析の中では基本的な方法であり、その考え方を理解することは、他の手法の理解にも通じるものと考えられる。ここでの分析には、統計学パッケージHALBAUを使用した[3][4][5]。実際の計算式の詳細は省略したが、興味ある読者は参考文献[3][5]などを参考にしてほしい。

＊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] 高木廣文（2007）：多変量解析を理解するために(1)：エストレーラ，1月号，pp.46-49.

[3] 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.

[4] 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.

[5] 高木廣文（2006）：HALBAU7によるデータ解析：シミック㈱.