日本行動計量学会 第25回大会発表論文抄録集, 1997

HALWINでの量的・質的変数の混在する場合の多変量解析について

高木　廣文（統計数理研究所）

【はじめに】

統計学パッケージHALBAUのWINDOWS版（HALBAU for WINDOW, 「HALWIN」と略す）では，多変量解析の多くの手法で，量的変数と質的変数の混在する場合について，簡単な設定で分析が行える。ここでは，各手法での変数の扱い，とくに質的変数のダミー変数への変換とそれに伴う検定や推定での処理方法について報告する。

【方法】

表１はHALWINに含まれる多変量解析の各手法の一覧と，変数の種類とその扱いを示したものである。重回帰分析などでは，説明変数に質的変数がある場合，総カテゴリ数と各カテゴリのデータの値を設定することで，（総カテゴリ数－１）個のダミー変数を分析のために作成する。この場合，第１番目のカテゴリに該当するダミー変数はない。分析に逆行列の計算が必要な手法は，すべて重回帰分析と同様の処理により，質的変数のダミー変数化を行なう。主成分分析では，質的変数の各カテゴリについて，反応があれば１，なければ０を与えて，分析を行なっている。

これらの方法では，年齢や身長のような量的変数であっても，同様の指定により質的変数として解析を行なうことが可能である。数量化理論では，同様の操作により，量的変数を質的変数に変換して分析が簡単に行なえる。

表１．HALWINでの多変量解析と変数の扱い

分析方法	量的変数	質的変数
1. 重回帰分析	○	ﾀﾞﾐｰ変数化
2. 数量化理論１類	順序尺度化	○
3. 主成分分析	○	ﾀﾞﾐｰ変数化
4. 正準相関分析	○	ﾀﾞﾐｰ変数化
5. 数量化理論３類	順序尺度化	○
6. 判別分析	○	ﾀﾞﾐｰ変数化
7. 数量化理論２類	順序尺度化	○
8. 多重ﾛｼﾞｽﾃｨｯｸﾓﾃﾞﾙ	○	ﾀﾞﾐｰ変数化
9. 指数ﾜｲﾌﾞﾙﾓﾃﾞﾙ	○	ﾀﾞﾐｰ変数化
10. 比例ﾊｻﾞｰﾄﾞﾓﾃﾞﾙ	○	ﾀﾞﾐｰ変数化
11. 条件付き多重ﾛｼﾞｽﾃｨｯｸﾓﾃﾞﾙ	○	ﾀﾞﾐｰ変数化
12. ﾛｼﾞｽﾃｨｯｸﾓﾃﾞﾙによる標準化	○	ﾀﾞﾐｰ変数化
13. 因子分析	○	ﾀﾞﾐｰ変数化
14. ｸﾗｽﾀｰ分析	○	----
15. 尺度構成と信頼性係数	○	○

（注）「変数」は基準変数がある場合には「説明変数」を，ない場合には分析に用いる変数を示す。
○：デフォルトで使用可能，またはもう一方の変数のようにも解析可能。

結局，HALWINでは表１の「14.クラスター分析」と「15.尺度構成と信頼性係数」では変数を変換しての解析ができないが，それ以外の分析ではすべて質的変数を含んだ解析が簡単にできる。以下では，代表的な方法として，重回帰分析と非線型モデルでの検定と推定での質的変数の扱いについて説明する。

１．重回帰分析での説明変数の有意性検定

重回帰分析でｋ個の説明変数がすでに含まれており，ダミー変数の個数を含めてｋ。個の変数があるとする。その重相関係数をR_kとし，そこに質的変数を追加した場合を考える。追加した変数の総カテゴリ数を

Cとすると，ダミー変数の個数は（C－1）個となり，変数の追加や削除はこのグループ単位で行なわれる。変数追加後の重相関係数をR_{k+1}とし，標本数をNとすると，検定のためには，

$$F_0=\frac{(N-k_c-C)}{(C-1)}\cdot\frac{(R_{k+1}^2-R_k^2)}{(1-R_{k+1}^2)}$$

を求め，F_0が自由度（C－1，N－k_c－C）のF分布に従うことを用いている。

2．非線型モデルでの説明変数の有意性検定

非線型モデルである「8.多重ロジスティックモデル」，「9.指数ワイブルモデル」，「10.比例ハザードモデル」，「11.条件付き多重ロジスティックモデル」，「12.ロジスティックモデルによる標準化」での質的変数の有意性検定の方法は，最尤法に基づく統一的な方法による。

基準変数をY，p個の説明変数をX_1，X_2，．．．，X_pとし，これらのN人のデータベクトルを$\mathbf{y}$，$\mathbf{x}_1$，$\mathbf{x}_2$，．．．，$\mathbf{x}_p$とする。このとき，説明変数にかける重みベクトルを$\boldsymbol{\beta}$とする。$\boldsymbol{\beta}=(\beta_1,\ \beta_2,\ \ldots,\ \beta_p)'$とすると，これらのモデルは，　$f(\mathbf{y})=g(\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_p,\ \boldsymbol{\beta})$ のように表せる。モデルの尤度関数を$L(\boldsymbol{\beta}\mid \mathbf{y},\ \mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_p)$とすると，尤度関数を最大にする$\boldsymbol{\beta}$の推定値$\mathbf{b}$を求めればよい。この場合，量的変数$X_i$の重み$b_i$の分散は，フィッシャー情報行列Iの逆行列Vの対角成分v_{ii}によって推定できる。

質的変数の場合，その総カテゴリ数をCとすると，ダミー変数は（C－1）個必要である。これらのダミー変数の重みを$\mathbf{b}_q=(b_{q1},\ b_{q2},\ \ldots,\ b_{q(C-1)})'$　とし，Vのこれらのダミー変数に対応する成分をv_{ij}，　(i, j=q1, ..., q(C-1)) とした場合，

$$\chi_0^2=\mathbf{b}_q'\begin{pmatrix} v_{q1q1} & v_{q1q2} & \cdots & v_{q1q(C-1)} \\ v_{q2q1} & v_{q2q2} & \cdots & v_{q2q(C-1)} \\ \vdots & \vdots & & \vdots \\ v_{q(C-1)q1} & \cdots & \cdots & v_{q(C-1)q(C-1)} \end{pmatrix}^{-1}\mathbf{b}_q$$

を求める。χ_0^2 が自由度（C－1）のχ^2 分布に従うことから，質的変数の有意性の検定を行なうことができる。「8.多重ロジスティックモデル」，「9.指数ワイブルモデル」，「10.比例ハザードモデル」，「11.条件付き多重ロジスティックモデル」，「12.ロジスティックモデルによる標準化」では，基準変数は生存・死亡などを表す変数であり，説明変数が質的な場合，その１番目のカテゴリの重みは０とし，それに対して他のカテゴリがどの程度のオッズ比（もしくはハザード比）をもつかを推定することが多い。カテゴリｉのオッズ比の点推定値は，$\exp(b_{q(i+1)})$で求められ，その$100(1-2\alpha)$％信頼区間は，　$\exp[b_{q(i+1)}\pm z_\alpha\sqrt{v_{q(i+1)q(i+1)}}]$　によって近似される。ここで，z_αは標準正規分布の上側α点である。

【おわりに】

ここでは，HALWINでの質的変数の扱いを限られた方法について説明したが，実際には量的変数と質的変数の混在する場合の多変量解析の各手法での変数の指定方法や，カテゴリの設定方法などについては説明していない。これらの点について，また分析結果の具体例などについては，当日報告する予定である。