数量化理論Ⅰ類における偏相関係数

中　山　　功

1　はじめに

多変量解析の重要な手法である数量化理論は，アンケート処理などの質的データの分析の場合には欠くことのできない手法である．戦後の日本で林知己夫氏を中心としたグループにより開発されてきた考え方［1］は，現在でも実践的な数多くの場面で中心的な分析手法として用いられてきており，特に最近では，コンピュータの発達とともに，各種の簡易ソフトにも組み込まれて，身近に容易に使用できる環境が整ってきている．筆者も講義やセミナーなどの演習の場面で，学生に課題と処理用プログラム［2］を与えて，分析を実行させることを試みているが，実践的なデータ処理を体験することによって，理論の意味を理解することも容易になっているように推察される．

この中で，予測モデルに相当するⅠ類とⅡ類では，回帰分析や判別分析と同様に，変量間の関係を考慮に入れながら，目的変量の予測と，その予測に重要な影響を及ぼす変量の考察（結果の目的変量に対して，いわゆる原因となる変量の考察）が中心的なものとなる．この原因あるいは要因となる変量の考察において重要な役割を果たすものが「偏相関係数」と呼ばれるものである．通常の「相関係数」が目的変量と原因となる説明変量との単純な相関の程度を測るのに対して，「偏相関係数」では，多変量の中で他の変量の影響を取り除いた残りの部分での真の相関の程度を数値化したことに相当して［3］，原因・結果の関連の程度をより厳密に測定・評価したものと考えることができる．したがって，この「偏相関係数」の数値を比較して，±1に近いものほど重要な要因となる変量であると判断するのは妥当な考え方である．ただし，回帰分析のときのように量的データから出発した分析では上述の考察には疑問の入り込む余地はあまりないと考えられるが，数量化理論では出発点に質的データがあり，「相関係数」でさえ，数量化の過程を経て初めて定義可能なもので，「偏相関係数」の定義と解釈にはより綿密な注意を払っていく必要があるのではないかと感じられる．

そこで，この論文では，偏相関係数の考察の原点に立ち戻って，各種数値計算例とともに，その解釈の妥当性を議論してみることにする．特に，回帰分析の場合には，偏相関係数の計算において3種類の一見異なるように見える導出法があるが，それらの同一性は解析的に証明されており，疑念の入り込む余地がないのは事実である．しかしながら，それに相対する数量化理論Ⅰ類においては，質的データを数量化する過程が必要とされて，上述の3種類の定義の一貫性に少なからず疑念の入り込む余地があるように感じられる．これに関して，解析的考察と数値計算とで議論を加えて，偏相関係数の妥当な解釈法を探ることを試みる．

以下では，まず回帰分析における偏相関係数の概略の考え方を示し，それに質的データを取り入れた数量化理論Ⅰ類の中での偏相関係数の位置づけを提示して，問題点を明確にする．次に，上述の3種類の定義を考察し直し，数量化理論Ⅰ類でのより妥当な解釈へと連なる道筋を探ることにする．さらに，数値計算例も交えて議論をすすめ，数量化理論Ⅰ類における偏相関

係数について，新たな解釈に導く考え方を紹介することを試みる．

2　回帰分析と偏相関係数

まず，数量化理論 I 類の議論に入る前に，回帰分析における偏相関係数の定義を振り返っておくことにする．目的変数を y として，p 個の説明変数 $x_1, x_2, \cdots, x_p$ と誤差項をあらわす ε を用いて，次のような予測モデルの式を考える．

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon. \tag{2.1}$$

この y と $x_1, x_2, \cdots, x_p$ に対する n 組の観測値が与えられたとき，(2.1) 式に代入して，誤差 ε が最も小さくなるように $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ を求めると，予測モデル式が求められることになる．これを行列記法で書き表すと以下のようになる．

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.2}$$

ここで，$\boldsymbol{y}$ は目的変数の n 個の観測値を縦に並べた列ベクトル，$\boldsymbol{X}$ は定数項の部分と p 個の説明変数の n 個の観測値を縦横に並べた行列，$\boldsymbol{\beta}$ は偏回帰係数 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ を縦に並べた列ベクトル，$\boldsymbol{\varepsilon}$ は各観測値に対する n 個の誤差を縦に並べた列ベクトルであり，具体的には次のように書ける．

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{2.3}$$

ここから，誤差の 2 乗の合計 $S_e = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$（$\boldsymbol{\varepsilon}'$ は転置行列を意味する）を最小にするように（最小 2 乗法），偏回帰係数ベクトル $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ を求めると，以下のように書ける．

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.4}$$

これから $\boldsymbol{y}$ に対する推測値 $\hat{\boldsymbol{y}}$ は

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{2.5}$$

で求められる．ただし，$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ である．また，このときの誤差（残差）は

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} \tag{2.6}$$

と計算できるので，残差平方和 S_e の値も

$$S_e = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})'(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}) = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\hat{\boldsymbol{y}} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} \tag{2.7}$$

より計算できる．これらより，F 値，決定係数，重相関係数などが計算されて，この回帰分析の当てはまりの良さなどが検証されることになる．

この回帰分析において，目的変数 y と説明変数 x_i との偏相関係数 $\tilde{r}_{yi}$ は次のように定義される．まず，x_i 以外の $p-1$ 個の説明変数 $x_1, x_2, \cdots, x_p$ を用いて，目的変数を y とした回帰分析を考

えたときの誤差の部分を u と定義し，同様に，x_i 以外の $p-1$ 個の説明変数 $x_1, x_2, \cdots, x_p$ を用いて，目的変数を x_i とした回帰分析を考えたときの誤差の部分を v と定義する．この u と v との相関係数を考えると，これが偏相関係数になる．すなわち，y と x_i において他の多変量 $x_1, x_2, \cdots, x_p$ に関連する部分を取り除いた残りの誤差 u と v との相関を考えることになって，真の相関の程度を数値化したことに相当していることが理解できる．

これを行列記法で書き表すと以下のようになる．x_i に対する n 個の観測値を縦に並べた列ベクトルを $\boldsymbol{x}_i$ と定義し，逆に，前述の行列 $\boldsymbol{X}$ において，x_i に対応する観測値の 1 列の部分を除去した行列を $\boldsymbol{X}_i$ と定義し，$\boldsymbol{H}_i=\boldsymbol{X}_i(\boldsymbol{X}'_i\boldsymbol{X}_i)^{-1}\boldsymbol{X}'_i$ とすると，

$$\boldsymbol{u}=\boldsymbol{y}-\boldsymbol{H}_i\boldsymbol{y},\quad \boldsymbol{v}=\boldsymbol{x}_i-\boldsymbol{H}_i\boldsymbol{x}_i, \tag{2.8}$$

$$\tilde{r}_{yi}=\frac{\boldsymbol{u}'\boldsymbol{v}}{\sqrt{\boldsymbol{u}'\boldsymbol{u}}\sqrt{\boldsymbol{v}'\boldsymbol{v}}} \tag{2.9}$$

である．ここで，p 個の説明変数 $x_1, x_2, \cdots, x_p$ 全体と y を含めた変数相互間の相関行列を考えたとき，その逆行列の各成分を r^{yy}, r^{yi}, r^{ij} などとすると，

$$\tilde{r}_{yi}=\frac{-r^{yi}}{\sqrt{r^{yy}r^{ii}}} \tag{2.10}$$

となることは，行列の計算の性質より解析的に証明できる［3］．

また，変数 x_i を説明変数として取り入れる前後での決定係数の増加分を考察して，偏相関係数に関連付けることも可能である．一般に x_i を含む p 個の説明変数全体での決定係数 R^2 は前述の行列記法と y の偏差平方和 $S_t=\boldsymbol{y}'\boldsymbol{y}-n\bar{y}^2$ を用いて，

$$R^2=1-\frac{S_e}{S_t}=\frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-n\bar{y}^2}{\boldsymbol{y}'\boldsymbol{y}-n\bar{y}^2} \tag{2.11}$$

と書き表される．この分子の回帰平方和 $S_r=\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-n\bar{y}^2$ の変数 x_i による増加分を考察すると，$\Delta S_r=\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}=\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}-\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}$ になるが，これと x_i を追加する前の残差平方和に相当する $S_e+\Delta S_r$ との比 $\Delta S_r/(S_e+\Delta S_r)$ を考える．これは正しく，残差の中で変数 x_i の追加により説明された部分の割合を示すことになり，変数 x_i の目的変数 y への真の寄与の程度を示すものと解釈され，偏決定係数 $\tilde{R}_i^2$ と定義されるものである．これは，(2.9), (2.10) の偏相関係数と関連付けられて，

$$\tilde{r}_{yi}^2=\tilde{R}_i^2=1-\frac{S_e}{S_e+\Delta S_r}=\frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}}=\frac{R^2-R_i^2}{1-R_i^2},\ R_i^2=\frac{\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}-n\bar{y}^2}{\boldsymbol{y}'\boldsymbol{y}-n\bar{y}^2} \tag{2.12}$$

になることが簡単な計算により示される．ここで，R_i^2 は変数 x_i を追加する以前の状態での決定係数を表しており，変数 x_i による決定係数の増加分 $R^2-R_i^2$ が偏相関係数の 2 乗に関連付けられることが理解できる．

これらの (2.9), (2.10), (2.12) が偏相関係数の計算における 3 つの主要な方法であり，特に (2.9) と (2.12) は偏相関の意味の解釈においても重要な役割を果たす表式であることは疑いないであろう．

3　数量化理論Ⅰ類と偏相関係数

前章での回帰分析における議論を説明変数が質的データである場合に拡張して，数量化理論Ⅰ類として考察する場合にも，ダミー変数を導入することにより，前述と同様な議論が可能にな

る．例えば，x_iが質的な変数（アイテムと呼ぶ）でm種類のデータ（カテゴリーと呼ぶ）の可能性があるとき，データを便宜上1, 2, …, mと考えると，これを$m-1$個のダミー変数で置き換えて議論することが可能になる［1］．このダミー変数の1–0データを (2.3) 式での $\boldsymbol{X}$ の行列要素と考えて，前述と同様の議論を展開すると，得られた偏回帰係数$\beta_1, \beta_2, \cdots$がカテゴリー数量に対応し（通常はこれらを平均0に規準化する），(2.1) の予測モデル式が求められることになる．また，そこで得られた各カテゴリー数量を元の質的データに当てはめて置き換えると，あたかも初めから量的データが存在するかのように考えることも可能である．例えば，この数量を (2.3) 式での $\boldsymbol{X}$ の行列要素と考えて，偏回帰係数ベクトル $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ を求めると，$\beta_0=\bar{y}, \beta_1=\beta_2=\cdots=1$となって，(2.1) 式は各カテゴリー数量の和で予測値を求めるという通常の数量化理論Ⅰ類としてのモデル式に帰着することになる．同様に，このカテゴリー数量を通常の量的データと同様に考えて相関係数を求めることも可能であり，その逆行列から，(2.10) 式より，偏相関係数を求めることも可能になることは容易に理解できる．

このように，得られたカテゴリー数量を元からの量的データと同様に取り扱ってよいものと考えると，(2.9) や (2.12) 式も同様に計算可能で，求められた3種類の偏相関係数の数値が全て一致することも明らかである．通常は，こうして求められた数値を偏相関係数として，変量の重要性などの要因分析にも利用されている場合が多いようである．しかしながら，例えば (2.9) 式で考えたとき，考察の元になった議論で，x_iを除いた$p-1$個の説明変数を用いて，目的変数をyとした回帰分析を考えて誤差$\boldsymbol{u}$を求める場合に，全変量の分析で求められたカテゴリー数量そのままを分析の数値として用いることには，偏相関の意味の解釈上からも大いに疑念の入り込む余地があると考えられる．むしろ，x_iを除いた$p-1$個の説明変数を用いて，最初からもう一度数量化理論Ⅰ類の分析を目的変数をyとして実行し直すべきで，得られた数量は当然に別物となり，誤差$\boldsymbol{u}$も異なった数値になるものと推察される．ただ，この場合は，最初は質的変量であるx_iを目的変数として誤差$\boldsymbol{v}$を求める定義が困難で，偏相関としての妥当な定義を見出すことは簡単ではない．

そこで，(2.12) 式の方に目を向けると，この場合は，x_iを質的変量としても，数量化理論Ⅰ類の議論により，決定係数であるR^2やR_i^2を妥当な解釈のままで求めることが可能である．したがって，そこから (2.12) 式を通じて求められた$\tilde{r}_{yi}$を新たに定義し直された偏相関係数として，解釈し直す方が適切ではないかと考えられる．ここで求められた数値は (2.9), (2.10) 式とは異なるものになると考えられるので，記法も$\tilde{r}'_{yi}$に変更しておくことにする．すなわち，

$$\tilde{r}'^2_{yi}=\tilde{R}_i^2=1-\frac{S_e}{S_e+\Delta S_r}=\frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}}=\frac{R^2-R_i^2}{1-R_i^2}, \qquad R_i^2=\frac{\boldsymbol{y}'\boldsymbol{H}_i\boldsymbol{y}-n\bar{y}^2}{\boldsymbol{y}'\boldsymbol{y}-n\bar{y}^2} \tag{3.1}$$

であり，ここでの$\boldsymbol{H}_i$などは通常の数量化理論Ⅰ類の手法でダミー変数により定義し直されているものとする．この場合，x_iを除いた分析での決定係数R_i^2の計算において，全変量での分析の場合のカテゴリー数量をデータとしてそのまま用いて計算したものより，ダミー変数に立ち戻って最適化を考え直していることから，決定係数の数値も最適なものになって，より大きな数値になるものと推察される．したがって，

$$\tilde{r}'^2_{yi} \le \tilde{r}^2_{yi} \tag{3.2}$$

となり，等号は上述の最適化の条件が同一のものとなる場合で，x_i以外の全質的変数がそれぞれダミー変数1つずつで書けている場合に対応する．すなわち，x_i以外の質的変数が全て2カ

テゴリーのときは，ここで新たに考え直した (3.1) の偏相関係数と従来のものとは一致するが，その他の場合は絶対値がより小さな数値として定義し直されることになる．表式 (3.1) における偏相関の意味としての解釈の妥当性からも，要因分析の場面では，こちらの数値を用いて議論し直した方が良いのではないかと考えられる．

以上の考察を実際のデータにより検証するため，次の 4 章では具体的な数値計算例を用いて議論をさらにすすめていくことにする．

4　数値計算例

ここでの数値計算例の出発点として，参考文献［1］で数量化理論Ⅰ類の分析で用いられている例を借用して，そこでの結果と比較・検討してみることにする．表4.1がその20人4変数のデータ例と，分析により得られた各カテゴリー数量及びそれを用いて計算された相関係数の行列である．

y 小遣い	x1 収入	x2 職業	x3 ﾏｽｺﾐ接触
10	1	1	1
8	1	3	2
18	2	1	1
30	3	4	2
15	1	4	1
12	2	3	2
22	2	1	1
12	1	3	2
25	3	2	3
8	1	1	2
20	3	1	3
10	2	3	1
14	2	1	2
6	1	3	1
28	3	2	3
8	2	4	3
10	1	2	1
25	2	1	3
28	3	1	1
28	2	2	2

アイテム	カテゴリー	数量
x1:収入	1	−6.85112
	2	0.48966
	3	8.80810
x2:職業	1	0.67388
	2	4.09846
	3	−4.54679
	4	0.31634
x3:ﾏｽｺﾐ接触	1	0.41383
	2	1.74344
	3	−3.10293

相関行列	y	x1	x2	x3
y	1	0.78378	0.57217	−0.28165
x1	0.78378	1	0.39502	−0.48794
x2	0.57217	0.39502	1	−0.40628
x3	−0.28165	−0.48794	−0.40628	1

表4.1　分析用データと数量，相関行列

この相関行列の逆行列から求められた偏相関係数と (3.1) で定義された偏相関係数の数値を比較したものが以下の表4.2である．

	(2.10)式	(3.1)式	比（%）
x1	0.77785	0.76649	98.54
x2	0.53819	0.53573	99.54
x3	0.36047	0.36003	99.88

表4.2　表4.1のデータ例での偏相関係数の比較

この表より，(3.1) で求められたものが，従来のものより少し小さな数値になっていることが確認できる．ただその差異はごくわずかなもので，数値の解釈の面で大きな影響を及ぼすまでには至っていないようである．次にカテゴリーを 2 にした場合の結果を確認するために，表4.1で x2 と x3のデータの 2 以上のカテゴリー値を全て 2 に書き直して分析し直した場合の偏相関係数の数値を比較したものが次の表4.3である．

	(2.10)式	(3.1)式	比（%）
x1	0.76824	0.76824	100
x2	0.08644	0.08509	98.45
x3	0.02424	0.02418	99.74

表4.3　表4.1のデータ例で x2 と x3を 2 カテゴリーにしたときの偏相関係数の比較

これより，x1については他の説明変数が全て 2 カテゴリーで，(3.2) が等号になる場合に相当していることが数値例でも確認できる．他の x2 と x3については，やはり99％前後のわずかな差異を生じているようである．

このように，実例で (3.2) 式の結果が確認できたが，(3.1) 式と従来の (2.10) 式の差は僅少であり，わざわざ新たな定義を持ち出すまでもなく，従来のままで，解釈に変更を及ぼすほどのものでもないようにも感じられる．別の例として，個体数と変量数の両方を大きくして，84 個体，9 説明変量（各 3 カテゴリー）のデータで分析した結果の数値を次の表4.4で示す（原データは省略する）．

	(2.10)式	(3.1)式	比（%）
x1	0.68790	0.67225	97.72
x2	0.62797	0.59223	94.31
x3	0.70731	0.67448	95.36
x4	0.57294	0.56348	98.35
x5	0.48470	0.42668	88.03
x6	0.84364	0.80709	95.67
x7	0.67971	0.65531	96.41
x8	0.61150	0.59187	96.79
x9	0.68020	0.62595	92.02

表4.4　84個体，9 説明変量のデータ例での偏相関係数の比較

この例でも，(3.1) 式と (2.10) 式との差はわずかであるが，偏相関係数の大きさの順序が逆転しているところもあるので, 要因分析の面で注意を喚起する意味はないとも言えないであろう．

他の各種の既存のデータ例を用いて計算を続行しているところであるが，現時点では，差異があるとしても 90 %を下回るものはほとんどなく，新たな (3.1) 式を特に強調すべき意味合いがあるかどうかは，計算の煩雑さとの兼ね合いで難しいところである．ただ，データ例によっては数値が大きく違って，やはり解釈に変更を加えるべき場合がないとは言い切れないので，今後とも注意して見ていくべきであろう．むしろシミュレーションなどの人工的な多数の例で比較・検討してみることも必要ではないかと感じられるが，それについては今後の課題としておく．

5　まとめ

前章までで明らかにされたように，数量化理論 I 類における偏相関係数の定義の再検討により，その問題点が明確になり，新たな定義の必然性が確認されたと考えられる．ただ，ここで紹介された新たな考え方では，既存の数値例に大きな変更を加えるほどのものはなく，実用的に考えて従来のものに完全に取って代わるところまでには至っていない点に，不十分さが残っていると感じられる．また，そもそも数量化理論 I 類における質的データの数量化の考え方が人工的な仮定に基づくもので，そこに敢えて量的な偏相関係数の考え方を持ち出す点に疑問が残るところもあろうが，ここはあくまで量的な分析の仮定の上で，どこまで自然な考え方で矛盾なく説明できるかという点に重点を置いている．

次に，ここでの新たな定義式 (3.1) には，従来のものより計算が煩雑になるという欠点はあるが，プログラム的にはそれほど難しいものはなく，一度プログラムとして組み込んでおけば，自然な解釈が可能な数値が得られることになり，その発展性には期待が持てるところである．さらにシミュレーションや解析的な考察などを積み重ねて，従来との差異の本質的な部分を追究していくべきところであろうが，これについては今後の課題として残しておく．

最後に，ここでの新たな偏相関係数の考え方は，当然，他の数量化理論 II 類などにも適用可能であり，また，他にも再検討されるべき数値の定義例が残されている可能性もある．もっと多くの分野で，応用範囲が広がっていくことを今後さらに期待し，注目していきたい．

参考文献

[1] 林知己夫監修，駒澤勉著，数量化理論とデータ処理，朝倉書店，1982.
[2] 駒澤勉，橋口捷久著，パソコン数量化分析，朝倉書店，1988.
[3] 田中豊，脇本和昌著，多変量統計解析法，現代数学社，1983.