数量化理論I類の方法とその適用例

Hayashi's Quantification Method Type I and its Application

ネットワーク情報学部　永添めぐみ　村上明日香　佐藤 創

School of Network and Information　Megumi NAGASOE, Asuka MURAKAMI, Hajime SATO

Keywords: Hayashi's Quantification Method, Statistical Data Analysis, Instant Messenger

まえがき

2006年度の授業科目「プロジェクト」において，インターネットにおけるメッセンジャーの使用状況に関するアンケート調査を行った．このとき，単純集計だけでなく数量化理論I類とよばれるデータ解析の方法も適用することを試みた．ここで用いた方法は大変に適用範囲が広いので，多くの人に知ってもらいたい．そためめに，この機会に分析結果の報告とその方法論の紹介をすることにした．

方法論の解説は佐藤，分析例の計算と結果については永添，村上が記す．

1　回帰分析

数量化理論I類の方法は多変量解析の方法の一つとしてデータ解析を行う人々の間ではよく知られている．しかし，本学ネットワーク情報学部では残念ながらよく知られているとは言えない（統計学より以前の線形代数の勉強が足りない）．

数量化I類は，回帰分析法の拡張と考えると容易に理解することができる．念のため，回帰分析の説明からはじめることにするので，重回帰分析に親しい方はこの節は読み飛ばしてよい．

まず準備として，平均，分散，共分散，相関係数，回帰直線などを簡潔に説明する．

1.1　1変量，平均，分散

例えば，n 人のクラス全員の走り幅跳びの記録

k	1	2	3	4	5	$\cdots$	n
x_k	4.2	5.1	4.4	5.3	4.9	$\cdots$	4.7

が与えられたとする．これらを**変量** X **の標本値**（サンプル，データ，観測値，測定値）であると考える．その**平均** m_x と**分散** s_{xx} は次のように定義される．

$$m_x = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad s_{xx} = \frac{1}{n}\sum_{k=1}^{n}(x_k - m_x)^2. \tag{1}$$

n を**標本数**という．分散について次の関係がある．

$$s_{xx} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - m_x^2. \tag{2}$$

分散 s_{xx} は s_x^2 と記されることも多く，$s_x = \sqrt{s_{xx}}$ を変量 X の**標準偏差**とよぶ．

平均 m_x はこのクラスの代表値，分散 s_{xx} または標準偏差 s_x はバラツキの程度を表している．

なお，各標本値 x_k に対して

$$u_k = \frac{x_k - m_x}{s_x}$$

を対応させることを**標準化**という．標準化された値 u_k の平均は 0，分散は 1 である．

1.2　2変量，単回帰，相関係数

例えば，n 人のクラス全員の走り幅跳びと走り高跳びの記録

k	1	2	3	4	5	$\cdots$	n
x_k	4.2	5.1	4.4	5.3	4.9	$\cdots$	4.7
y_k	83	77	86	75	95	$\cdots$	82

が与えられたとする．2変量 X, Y の標本値の間の**共分散** s_{xy} と**相関係数** r_{xy} を次のように定義する．

$$s_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - m_x)(y_k - m_y), \quad r_{xy} = \frac{s_{xy}}{s_x\, s_y}. \tag{3}$$

(2) と同様に，$s_{xy} = \frac{1}{n}\sum_{k=1}^{n} x_k\, y_k - m_x\, m_y$ であり，とくに，X, Y が標準化されて変量 U, V で表されるとき，

$$r_{xy} = r_{uv} = s_{uv} = \frac{1}{n}\sum_{k=1}^{n} u_k\, v_k$$

である．相関係数には

$$-1 \le r_{xy} \le 1 \tag{4}$$

の性質があり，r_{xy} は2変量 X, Y の間の**相関関係**（線形関係）の強さを表す．$r_{xy} = 0$ のとき，2変量 X, Y は

統計的に相関がないという（X, Y が**互いに独立**ならば $r_{xy}=0$ であるが，逆は真ならず）．$r_{xy}>0$ のときは**正の相関**がある，$r_{xy}<0$ のときは**負の相関**がある，といい，絶対値 $|r_{xy}|$ の大きいほど**相関が強い**，という．

とくに，$r_{xy}=\pm 1$ のとき2変量 X, Y の間に線形関係

$$Y = aX + b$$

がある（定数 a の符号は r_{xy} の符号と同じ）．

一般に，標本値 x_k, y_k の間に近似的に線形関係

$$y_k = a\,x_k + b + e_k \qquad (k=1,2,\cdots,n) \tag{5}$$

が成り立つものと想定し，誤差 e_k の2乗和最小の条件

$$\sum_{k=1}^{n} (e_k)^2 \;\to\; \min \qquad (\text{最小化}) \tag{6}$$

のもとで線形式（直線の式）

$$y = a\,x + b \tag{7}$$

を求める作業を**回帰分析**といい，式(7)を**回帰直線**とよぶ．

条件(6)により**回帰係数** a, b を求める方法を**最小2乗法**とよぶ．結果として a, b は次のように計算すればよい．

$$a = \frac{s_{xy}}{s_{xx}}, \qquad b = m_y - a\,m_x. \tag{8}$$

相関係数との間に，$a = r_{xy}\left(\dfrac{s_y}{s_x}\right)$ という関係がある．

回帰直線(7)は，x の値が与えられたときに y の値を予測するための**予測式**と考えることができる．このとき実測値 y に対して $e = y-(a\,x+b)$ を**予測誤差**または**残差**とよぶ．標本値に関する誤差 e_k の平均は0，分散は

$$\frac{1}{n}\sum_{k=1}^{n} (e_k)^2 = s_{yy}(1-r_{xy}^2)$$

で与えられるから，相関係数 r_{xy} の絶対値が小さいときの予測はあまり正確ではなく，目安として $|r_{xy}|>0.7$ のとき，「2変量 X, Y に関する**線形モデル** $Y = aX+b$ は**意味を持つ**」と考える．

参考　式(8)は次のように導かれる．誤差の2乗和

$$\textstyle\sum_{k=1}^{n} (e_k)^2 = \sum_{k=1}^{n} (a\,x_k + b - y_k)^2$$

は2変数 a, b の関数であり，極値（最小値）をとる点は2つの偏微分係数が0となる a, b である．したがって，

$$\begin{cases} \displaystyle\sum_{k=1}^{n} (a\,x_k + b - y_k)\,x_k = 0, \\ \displaystyle\sum_{k=1}^{n} (a\,x_k + b - y_k) = 0 \end{cases} \tag{9}$$

である．(9)の第2式より $b = m_y - a\,m_x$ が得られ，これを第1式の左辺に代入すると

$$\sum_{k=1}^{n} \Big(a(x_k-m_x)^2 - (x_k-m_x)(y_k-m_y) \Big) = n(a\,s_{xx} - s_{xy})$$

となる．したがって，$a = \dfrac{s_{xy}}{s_{xx}}$ が得られる．

注意　式(7)は，厳密には，変量 Y に関する変量 X による回帰直線という．変量 X に関する変量 Y による回帰直線

$$x = a'\,y + b'$$

を，単に式(7)を変形して $a' = 1/a$, $b' = -b/a$ と求めてはならない．正しくは次の通りである．

$$a' = \frac{s_{xy}}{s_{yy}} = r_{xy}\left(\frac{s_x}{s_y}\right) = a\left(\frac{s_x}{s_y}\right)^2, \quad b' = m_x - a'\,m_y.$$

1.3　3変量，重回帰，重相関，偏相関

例えば，n 人のクラス全員の走り幅跳びと走り高跳びの記録と身長

k	1	2	3	4	5	$\cdots$	n
x_k	4.2	5.1	4.4	5.3	4.9	$\cdots$	4.7
y_k	83	77	86	75	95	$\cdots$	82
z_k	133	146	155	137	140	$\cdots$	153

が与えられたとする．このデータを3変量 X, Y, Z の標本値と考えて，次のような**線形モデル**

$$Z = a\,X + b\,Y + c. \tag{10}$$

を当てはめる作業を**重回帰分析**という．X, Y を**説明変量**，Z を**被説明変量**とよぶ．習慣的に説明変量が1個のとき単回帰，複数個のとき重回帰と区別するが，すべて回帰分析としてもよい．

単回帰のときと同様に，標本値 x_k, y_k, z_k の間に近似的な線形関係

$$z_k = a\,x_k + b\,y_k + c + e_k \quad (k=1,2,\cdots,n) \tag{11}$$

を想定し，誤差 e_k の2乗和 $\sum_{k=1}^{n} (e_k)^2$ が最小となる**回帰平面**

$$z = a\,x + b\,y + c \tag{12}$$

の回帰係数 a, b, c を最小2乗法によって求める．

単回帰のときの式(9)はこの場合，

$$\begin{cases} \displaystyle\sum_{k=1}^{n} (a\,x_k + b\,y_k + c - z_k)\,x_k = 0, \\ \displaystyle\sum_{k=1}^{n} (a\,x_k + b\,y_k + c - z_k)\,y_k = 0, \\ \displaystyle\sum_{k=1}^{n} (a\,x_k + b\,y_k + c - z_k) = 0 \end{cases} \tag{13}$$

のように拡張され，第3式より，

$$a\,m_x + b\,m_y + c = m_z \tag{14}$$

が得られ，$c = m_z - a\,m_x - b\,m_y$ を第 1，2 式に代入すると a, b に関する連立方程式

$$\begin{cases} a\,s_{xx} + b\,s_{xy} = s_{xz}, \\ a\,s_{xy} + b\,s_{yy} = s_{yz} \end{cases} \tag{15}$$

が得られる．この関係を行列で表すと，

$$\begin{bmatrix} s_{xx} & s_{xy} \\ s_{xy} & s_{yy} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} s_{xz} \\ s_{yz} \end{bmatrix} \tag{16}$$

となる．式 (15)，(16) は説明変量が一般に m 個のときの連立方程式を示唆している．これは**正規方程式**とよばれる．また，相関係数を成分とする行列

$$R = \begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix} \tag{17}$$

は**相関行列**とよばれる．

式 (15) または (16) と (14) から定まる a, b, c により z_k の予測値 $\hat{z}_k = a\,x_k + b\,y_k + c$ が求まるが，z_k と $\hat{z}_k$ の間の相関係数を変量 Z に対する X, Y の**重相関係数**とよび，$r_{z,xy}$ で表す．相関行列 R の逆行列の各成分を R_{ij} で表すことにすれば，

$$r_{z,xy} = \sqrt{1 - \frac{1}{R_{zz}}} \tag{18}$$

となる（行列 R の要素 r_{ij} の余因子を R_{ij} で表すことがあるが，ここではそうでないことに注意）．

重相関係数 $r_{z,xy}$ は常に 0 以上の値を取るが，目安として $r_{z,xy} > 0.7$ のとき考察している線形モデルに意味があり，予測が有効になる．(なお，単回帰のときの重相関係数 $r_{z,x}$ は単相関係数 r_{zx} の絶対値になる．)

さて，Z は X, Y によって説明され，X は Y によって説明されるから，単純な相関係数 r_{zx} は変量 Y の影響が含まれている．そこで，純粋に Z と X の相関関係を調べるには，Z，X それぞれから X 以外の説明変量で説明できる部分を取り去った量を考えて，その 2 つの量の相関を考える必要がある．この場合，回帰式

$$z = a_z\,y + b_z, \quad x = a_x\,y + b_x$$

を求めて，2 つの量

$$z(y)_k = z_k - (a_z\,y_k + b_z), \quad x(y)_k = x_k - (a_x\,y_k + b_x)$$

に関する相関係数を考える．これを Z と X の間の**偏相関係数**とよび，$r_{zx,y}$ で表す．偏相関係数 $r_{zy,x}$ についても同様である．それらは次の関係によって求めることができる（これは強力な公式である．機会があればその理由を理解しておくとよい）．

$$r_{zx,y} = -\frac{R_{zx}}{\sqrt{R_{zz}\,R_{xx}}}, \quad r_{zy,x} = -\frac{R_{zy}}{\sqrt{R_{zz}\,R_{yy}}}. \tag{19}$$

相関係数 r_{zx}, r_{zy} を偏相関係数と区別して**単相関係数**とよぶ．単相関が弱くても偏相関係が強いこともあるので，重回帰分析を行う意味がある．

2　数量化 I 類の方法

数量化 I 類とよばれるデータ解析の方法は，被説明変量は**数量的**であるが，説明変量がカテゴリーへの分類で表される**カテゴリカル**な場合に，回帰分析の方法を拡張したものである（I 類以外の数量化法の説明は省略）．

2.1　カテゴリー値と数量化

例えば，n 人のクラス全員の血液型と性別と走り幅跳びの記録

k	1	2	3	4	5	$\cdots$	n
x_k	A	O	B	AB	A	$\cdots$	B
y_k	男	女	女	男	男	$\cdots$	女
z_k	4.2	5.1	4.4	5.3	4.9	$\cdots$	4.7

が与えられたとする．

一般に，変量 X は s 個のカテゴリー $X_1, X_2, \cdots, X_s$ への分類，変量 Y は t 個のカテゴリー $Y_1, Y_2, \cdots, Y_t$ への分類とする．例えば血液型 A, B, O, AB をそれぞれ X_1, X_2, X_3, X_4 に対応させ，血液型 O であることを $x_k = (0,0,1,0)$ で表す．また，性別の男，女をそれぞれ Y_1, Y_2 に対応させ，例えば女であることを $y_k = (0,1)$ で表す．

上のデータを $s = 4, t = 2$ の場合の 3 変量

$$X = (X_1, \cdots, X_s), \quad Y = (Y_1, \cdots, Y_t), \quad Z$$

の標本値と考えて，次のようなモデル

$$Z = a \cdot X + b \cdot Y + c. \tag{20}$$

を当てはめる作業を変量 X, Y の**数量化**という．ここに，$a = (a_1, \cdots, a_s)$, $b = (b_1, \cdots, b_t)$ を数ベクトルとし，

$$a \cdot X = \sum_{i=1}^{s} a_i\,X_i, \qquad b \cdot Y = \sum_{j=1}^{t} b_j\,Y_j$$

とする（$a \cdot X$ などは**ベクトルの内積**に相当する）．

標本値 $x_k = (x_{k1}, x_{k2}, x_{k3}, x_{k4})$, $y_k = (y_{k1}, y_{k2})$, z_k の間に線形関係

$$z_k = a \cdot x_k + b \cdot y_k + c + e_k \quad (k = 1, 2, \cdots, n) \tag{21}$$

を想定し，誤差 e_k の 2 乗和 $\sum_{k=1}^{n} (e_k)^2$ が最小となる**線形回帰式**

$$z = a \cdot x + b \cdot y + c \tag{22}$$

の係数 $a = (a_1, \cdots, a_s), b = (b_1, \cdots, b_t), c$ を最小 2 乗法によって求める．

ただし，関係 $\sum_{i=1}^{s} x_{ki} = 1$, $\sum_{j=1}^{t} y_{kj} = 1$ があるから，このままでは a, b, c は一意に定まらない．そこで，

$$a \cdot f_x = 0, \qquad b \cdot f_y = 0 \tag{23}$$

という制約をおく．ここに，

$$f_i^x = \sum_{k=1}^{n} x_{ki}, \qquad f_j^y = \sum_{k=1}^{n} y_{kj}, \\ f_x = (f_1^x, \cdots, f_s^x), \quad f_y = (f_1^y, \cdots, f_t^y) \tag{24}$$

とおく．例えば，f_i^x はカテゴリー X_i に属する標本の頻度である．さらに，

$$f_{ij}^{xy} = f_{ji}^{yx} = \sum_{k=1}^{n} x_{ki}\, y_{kj} \text{ (クロス頻度)}, \\ g_i^x = \sum_{k=1}^{n} z_k\, x_{ki}, \quad g_j^y = \sum_{k=1}^{n} z_k\, y_{kj}, \quad g = \sum_{k=1}^{n} z_k \tag{25}$$

とおけば，式 (13) に相当する式は

$$\begin{cases} a_i f_i^x + \sum_{j=1}^{t} b_j f_{ij}^{xy} + c f_i^x = g_i^x \quad (i = 1, \cdots, s), \\ \sum_{i=1}^{s} a_i f_{ji}^{yx} + b_j f_j^y + c f_j^y = g_j^y \quad (j = 1, \cdots, t), \\ \sum_{i=1}^{s} a_i f_i^x + \sum_{j=1}^{t} b_j f_j^y + cn = g, \\ \sum_{i=1}^{s} a_i f_i^x = 0, \\ \sum_{j=1}^{t} b_j f_j^y = 0 \end{cases} \tag{26}$$

となる．未知数の個数 $s+t+1$ より等式の個数が 2 個多いが，これは最初の s 個の式の辺々の和，およびその次の t 個の式の辺々の和が，ともに最後から 3 番目の式と一致することから生じている．そこで，最初の s 個の式，およびその次の t 個の式の中からそれぞれ 1 個の式を除去すればよい．

また，最後の 3 つの式から，$c = g/n = m_z$ が得られる．

この連立方程式を解いて得られる値 $a = (a_1, \cdots, a_s)$, $b = (b_1, \cdots, b_t)$ は各変量の**カテゴリー値**とよばれる．その結果，分類で与えられる変量 X, Y の標本値を

$$x_k = \sum_{i=1}^{s} a_i\, x_{ki}, \qquad y_k = \sum_{j=1}^{t} b_j\, y_{kj} \tag{27}$$

という数量に置き換えることができる（これが**数量化**の意味である）．

簡便的に，各変量のカテゴリー値の**レンジ** (range)

$$r_x = \max_{1 \le i \le s} a_i - \min_{1 \le i \le s} a_i, \quad r_y = \max_{1 \le j \le t} b_j - \min_{1 \le j \le t} b_j \tag{28}$$

が相関係数の代わりに用いられることが多いが，レンジは少数の偏りに影響されやすいので注意が必要である．

2.2 数量化にもとづく回帰分析

数量化にもとづく変量に関する分散，共分散は，$m_x = 0$, $m_y = 0$ であるから次のように計算される．

$$s_{xx} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 = \frac{1}{n}\sum_{i=1}^{s} a_i^2 f_i^x, \quad s_{yy} = \frac{1}{n}\sum_{j=1}^{s} b_j^2 f_j^y, \\ s_{xy} = s_{yx} = \frac{1}{n}\sum_{k=1}^{n} x_k\, y_k = \frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{t} a_i\, b_j\, f_{ij}^{xy}, \\ s_{xz} = s_{zx} = \frac{1}{n}\sum_{i=1}^{s} a_i\, g_i^x, \quad s_{yz} = s_{zy} = \frac{1}{n}\sum_{j=1}^{t} b_j\, g_j^y.$$

その後は通常の回帰分析（1.3 節）と同じで，単相関を求め，相関行列 (17) から，式 (18) により重相関係数，式 (19) により偏相関係数を計算することができる．

2.3 拡張

数量化 I 類の方法に関して，説明変量の個数を 1 以上任意に拡張することは容易であろう．

さらに，カテゴリーへの属し方が 0,1 ではなく確率的である場合に拡張することも容易である．

数量的な変量も**非線形性**が予想される場合には，適当な区間を設定し，値をその属する区間で表現するカテゴリカルな変量として扱えば，カテゴリー値の並びから非線形関係を知ることができる．

また，数量的な変量とカテゴリカルな変量の混在する**混合型モデル**に拡張することもできる．

このように回帰分析法を数量化の方法によってを拡張すれば，その適用範囲は格段に広がり，統計的なデータ解析の有力な“武器”が提供されることになるであろう．

3 数量化を行うにあたって

3.1 アンケート調査

私たちは 2006 年度「プロジェクト」を,「コミュニケーションツールの可能性を探る」というテーマで取り組んだ．従来のツールである電話・メールにかわり，近年新たに登場したメッセンジャーに注目し，それを自分たちで試作しつつ，未来のコミュニケーションのあり方を展望する企画である．**メッセンジャー**とは，同じシステム環境にある仲間どうしがチャットやファイル転送などを容易に行なうことを可能にした「インスタントメッセンジャー」(IM) と呼ばれるソフトウエアのことである．

プロジェクトチームは 2 手に分かれ，実装班が C++ 言語によりメッセンジャーの作成に取り組む一方で，調査・分析班はメッセンジャーがどのような状況で使用されているかを，電話・メールと対比しながら明らかにし

ようとした．その目的で，利用の実態に関する小規模なアンケート調査を試みることになった．

アンケートの実施後，今後の展開を調査・考察するグループと，アンケートを集計・分析するグループに分かれた．本稿は，その後者の行った努力の記録である．

アンケートの概要は以下の通りである．

調査対象　専修大学学生（1年生）

学部	男	女	不明	計
経営	73	25	1	99
ネットワーク情報	168	65	1	234
計	241	90	2	333

調査期間　2006年10月26日，27日

調査方法　先生方の協力で授業の開始前20分を頂き，回答用紙(A4片面3枚)を配布，回収した．設問数19，選択肢での回答と自由記述．

3.2　単純集計

得られたアンケートデータはExcelに入力し，それぞれの回答数を集計し，その結果をグラフで表示するなどの統計分析を行った．集計結果により，電話，メールの使用率はほぼ100％であるのに対して，メッセンジャーの使用率は全体で約50％であること，また経営学部1年生よりネットワーク情報学部1年生の方が圧倒的に多いことがわかった．

Q11　使ったことがあるものはどれですか．

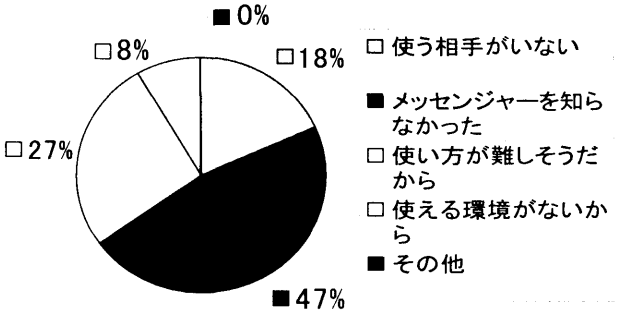
学部	電話	メール	メッセンジャー
経営	97％ (96)	97％ (96)	21％ (21)
ネット	98％ (230)	100％ (233)	62％ (145)
合計	98％ (326)	99％ (329)	50％ (166)

それぞれ使ったことがあると回答した人数の割合
(小数点以下四捨五入、カッコ内は人数)

集計結果をグラフ化した例を以下に示す．

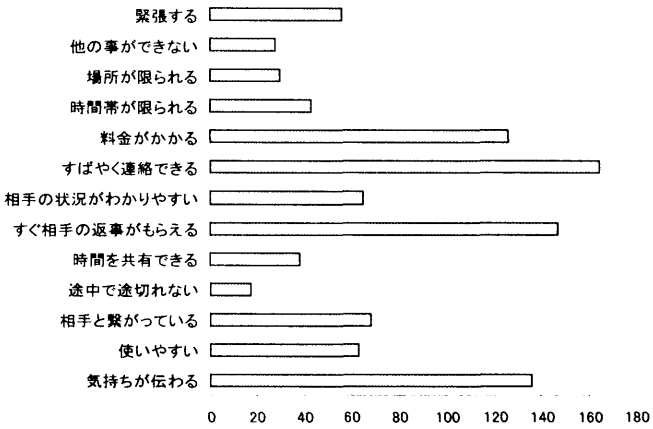
Q17は，Q11でメッセンジャーを使ったことがないと答えた人に，使わない理由を訊ねる質問である．

Q17　メッセンジャーを使っていない理由は？



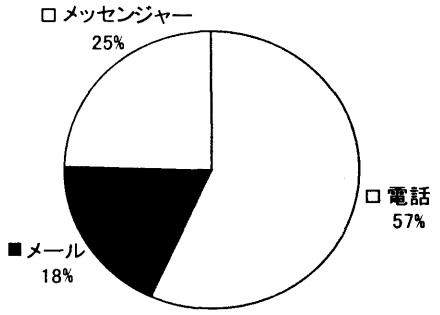
Q8では，テキストコミュニケーションと対比させる意味で音声コミュニケーションのイメージについて質問した．

Q8　音声コミュニケーションのイメージは？



Q10は「おしゃべり」をするときにどんな手段を使う人が多いかを知るための質問である．

Q10　おしゃべりをしたときに使う手段は？



その他の結果の紹介は省略するが，**単純集計**で得られたものはほぼ想定される範囲内であり，際立った特徴を見出すことはできそうもなかった．学生メンバーだけではせっかく入手したデータをどの様に生かせばよいか，判らないでいた．

しかし，本プロジェクトでは予め前期後半，毎週の授業で，担当教員(佐藤)により「回帰分析と数量化Ⅰ類」の15分間連続講座（4回）を行っていた．そこで今回のアンケート調査において「数量化手法の導入」を試みることになったのである．

4　数量化理論Ⅰ類の適用

4.1　データの説明

分析目標は,「メッセンジャーの個人的使用頻度がその人の電話・メールの使用頻度とどのような関係にあるか」を明らかにすることで，そのために数量化Ⅰ類を適用する．

対象データについて説明する．このデータは，次の質問に対する回答によって得られたものである．

Q14 各ツールをどのくらいの頻度で使用しますか．

電話 (1 日)	0 分　1〜10 分　30 分以下　それ以上
メール (1 日)	0 通　1〜5 通　10 通以下　20 通以下　それ以上
メッセンジャー (1 週間)	普段全く使わない　1〜2 回　3〜4 回　5〜6 回　毎日

データは，回答者が該当項目を選択することによって得られる．選択肢は量的な順序をもつが，カテゴリカルな変量であり，このままでは回帰分析の対象にはできない．

変量を次のように記号化して 2 節と関連付けることにする．各個人の電話・メールの利用頻度を表す説明変量をそれぞれ X,Y とし，メッセンジャーの使用頻度を表す被説明変量を Z とする．さらに，性別・学部を表す説明変量を追加してそれぞれ U,V とする．

すなわち，変量 Z を変量 X,Y,U,V によって統計的に説明するわけである．説明変量 X,Y,U,V の各カテゴリーと，カテゴリー値を表す変数を次のように定める．

参考までに各カテゴリーに属する人数を併記した．なお，意図していない回答を含むデータを排除したため，標本数 n は 270 に減少している。また，性別不明の 2 名は過半数を占める男性に含めることにした．

X	a	電話 (1 日)	人数
X_1	a_1	使わない	2
X_2	a_2	0 分	51
X_3	a_3	1〜10 分	142
X_4	a_4	11〜30 分	42
X_5	a_5	それ以上	33

カテゴリー「使わない」は，Q11 で「電話を使わない」と回答したことを表し，カテゴリー「0 分」は電話を使うが最近は使っていないことを表す．

Y	b	メール (1 日)	人数
Y_1	b_1	0 通	14
Y_2	b_2	1〜5 通	90
Y_3	b_3	6〜10 通	85
Y_4	b_4	11〜20 通	52
Y_5	b_5	それ以上	29
U	p	性別	人数
U_1	p_1	男	198
U_2	p_2	女	72
V	q	学部	人数
V_1	q_1	経営	84
V_2	q_2	ネットワーク情報	186

数量化 I 類の方法を適用するには，非説明変量 Z は量的でなければならない．そこで，メッセンジャーの使用頻度に関しては便宜的に，Q11 で「メッセンジャーを使ったことがない」と回答した場合 (0* で示す) に頻度 0，それ以外で Q14 で「普段全く使わない」と回答した場合 (00 で示す) に頻度 1，以下順に，頻度 2, 3, 4, 5 と定め，それらを量として扱うことにした．

人数の分布は次のようになり，変量 Z の平均と標準偏差は $m_z = 1.477, s_z = 1.650$ となる．

	0*	00	1-2	3-4	5-6	毎日	
Z の値	0	1	2	3	4	5	計
人数	111	44	58	19	8	30	270

なお，本来ならこの説明変量 Z もカテゴリカルなので，**数量化 II 類**が適用されるべきであるが，簡便法として自ら変量 Z の数量化を行ったわけである．

4.2 計算プロセス

アンケートの結果より得られたデータは下記の表になる．これは，2.1 節の最初の表を転置したものに当たる．すなわち，Z の欄の数値は量を表し，それ以外はカテゴリーを選択する番号である．

No	X	Y	U	V	Z
1	2	2	1	2	5
2	1	2	2	1	1
3	1	1	2	1	0
4	3	4	1	2	4
5	3	5	2	2	1
⋮	⋮	⋮	⋮	⋮	⋮
270	2	3	1	2	0

参考までに，今記述している分析の前に行った結果を記しておきたい．最初は，メッセンジャーの使用頻度に対する電話・メールの使用頻度が与える影響を調べるために，電話 (X)，メール (Y)，メッセンジャー (Z) の 3 変量だけで分析を行っていた．そのとき得られた結果は，重相関係数が

$$r_{z,xy} = 0.202\cdots$$

であるに過ぎず，単相関係数とレンジは

変量	単相関	レンジ
X(電話)	0.143	1.517
Y(メール)	0.141	0.735

であった（偏相関係数は計算してない）．このように，X,Y の 2 変量では説明力が弱かったため，性別 (U) と学部 (V) の 2 つの説明変量を追加したのである．

まず，270 個のデータに関するクロス集計を行う．これには Excel のクロス集計機能であるピボットテーブル

を用いた．得られた数値は式 (24), (25) で定義されたもので，式 (26) の係数が得られる．それを以下の表に示す．

a_1	a_2	a_3	a_4	a_5	b_1	b_2	b_3	b_4	b_5	p_1	p_2	q_1	q_2	c	1
2	0	0	0	0	0	1	0	0	1	2	0	0	2	2	5
0	51	0	0	0	11	28	7	3	2	38	13	13	38	51	73
0	0	142	0	0	3	51	53	29	6	110	32	41	101	142	215
0	0	0	42	0	0	7	12	14	9	32	10	14	28	42	75
0	0	0	0	33	0	3	13	6	11	16	17	16	17	33	31
0	11	3	0	0	14	0	0	0	0	12	2	4	10	14	13
1	28	51	7	3	0	90	0	0	0	74	16	22	68	90	154
0	7	53	12	13	0	0	85	0	0	61	24	28	57	85	132
0	3	29	14	6	0	0	0	52	0	29	23	20	32	52	67
1	2	6	9	11	0	0	0	0	29	22	7	10	19	29	33
2	38	110	32	16	12	74	61	29	22	198	0	63	135	198	331
0	13	32	10	17	2	16	24	23	7	0	72	21	51	72	68
0	13	41	14	16	4	22	28	20	10	63	21	84	0	84	37
2	38	101	28	17	10	68	57	32	19	135	51	0	186	186	362
2	51	142	42	33	14	90	85	52	29	198	72	84	186	270	399
2	51	142	42	33	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	14	90	85	52	29	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	198	72	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	84	186	0	0

例えば，この表の第 1 行目は

$$2a_1 + b_2 + b_5 + 2p_1 + 2q_2 + c = 5$$

という等式を表す．この連立方程式の変数の個数は 15(= 5 + 5 + 2 + 2 + 1) であるが，式の数は 19 個である．線形従属性を除くために例えば，第 1, 6, 11, 13 式を除けば独立な 15 個の式が得られる．こうすれば，通常の解法で解を求めることができる (2.1 節参照)．

なお，最後の 5 式より，$c = 399/270 = 1.477\,(= m_z)$ が求まる．

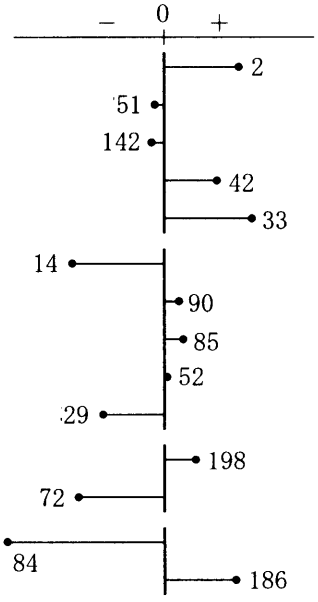
参考　Excel 関数による連立方程式の解法

まず係数行列の逆行列を関数 MINVERSE を用いて計算する．それには，「MINVERSE(範囲)」という式によって行列が格納されているセルの範囲を指定する．このとき，計算結果を入れるセルの範囲を指定する必要がある．その簡便法は，まず MINVERSE の式を 1 つのセルに入力し，そこを始点として必要な範囲を選択する．次に数式バーをクリックし，Ctrl キー+Shift キー+Enter キーを押せば，指定した範囲に逆行列が入る（プロジェクトのリーダー高野智弘君より教わる）．

解を逆行列と定数ベクトルの積として求めるために，MMULT 関数を使用する．これは 2 つの行列の積を「MMULT(範囲 1, 範囲 2)」という式によって求める関数で，計算結果を入れる範囲を指定するには上記と同じ手順で行えばよい．

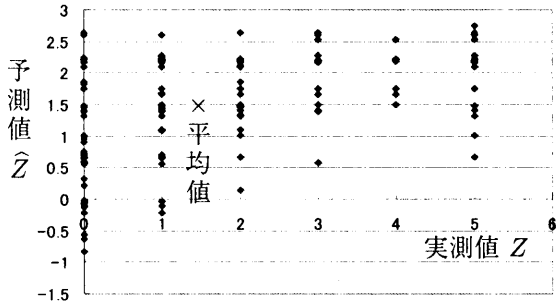
各カテゴリー値は以下の通りであった．視覚的にわかりやすいように，数値を棒グラフで表した図を添えた．

変量		カテゴリー値
X	a_1	0.495
	a_2	−0.064
	a_3	−0.085
	a_4	0.348
	a_5	0.581
Y	b_1	−0.615
	b_2	0.098
	b_3	0.125
	b_4	0.018
	b_5	−0.409
U	p_1	0.208
	p_2	−0.573
V	q_1	−1.049
	q_2	0.474
定数	c	1.477



(数字は人数を表す)

これらのカテゴリー値を用いて各変量を数量化すると，Z の予測値 $\widehat{Z} = a \cdot X + b \cdot Y + p \cdot U + q \cdot V + c$ が得られる．点 $(Z, \widehat{Z})$ を平面上にプロットした図を示す．一見して予測があまり正確でないことがわかる．



2.2 節 に述べた方法により分散・共分散を計算し，1.3 節の手順で単相関行列 R, その逆行列 R^{-1} から重相関係数や偏相関行列 P を計算する．

$$R = \begin{bmatrix} 1.000 & -0.108 & 0.009 & -0.022 & 0.075 \\ -0.108 & 1.000 & -0.035 & 0.018 & 0.120 \\ 0.009 & -0.035 & 1.000 & -0.025 & 0.195 \\ -0.022 & 0.018 & -0.025 & 1.000 & 0.422 \\ 0.075 & 0.120 & 0.195 & 0.422 & 1.000 \end{bmatrix}$$

行列 R の 5 行目に単相関係数 $r_{zx}, r_{zy}, r_{zu}, r_{zv}$ が並ぶ．(r_{zx}, r_{zy} が 3 変量モデルと異なるのは数量化のため)

$$R^{-1} = \begin{bmatrix} 1.024 & 0.125 & 0.021 & 0.075 & -0.127 \\ 0.125 & 1.035 & 0.070 & 0.057 & -0.171 \\ 0.021 & 0.070 & 1.059 & 0.143 & -0.277 \\ 0.075 & 0.057 & 0.143 & 1.243 & -0.565 \\ -0.127 & -0.171 & -0.277 & -0.565 & 1.323 \end{bmatrix}$$

逆行列 R^{-1} から重相関係数は次のように求まる.

$$r_{z,xyuv} = \sqrt{1 - 1/1.323} = \mathbf{0.494}\cdots.$$

偏相関行列 P は次のようになり，5 行目に偏相関係数 $r_{zx,yuv}, r_{zy,xuv}, r_{zu,xyv}, r_{zv,xyu}$ が並ぶ.

$$P = \begin{bmatrix} -1.000 & -0.122 & -0.021 & -0.067 & 0.110 \\ -0.122 & -1.000 & -0.067 & -0.051 & 0.147 \\ -0.021 & -0.067 & -1.000 & -0.125 & 0.234 \\ -0.067 & -0.051 & -0.125 & -1.000 & 0.441 \\ 0.110 & 0.147 & 0.234 & 0.441 & -1.000 \end{bmatrix}$$

これらの計算結果を表にまとめて示す.

変量	単相関	偏相関	レンジ	順位
X(電話)	0.075	0.110	0.581	4
Y(メール)	0.120	0.147	0.740	3
U(性別)	0.195	0.234	0.781	2
V(学部)	0.422	0.441	1.523	1

4.3 結果の考察

カテゴリー値に関して 変量 X のカテゴリー値 $a_1,\cdots,a_5$ を見ると興味深いことがわかる．電話をまったく使わない人と最も使う人のカテゴリー値が大きく，電話を使う頻度が中程度の人のカテゴリー値は小さい．これより，電話を使わない人はメッセンジャーを電話代わりに使用し，電話を頻繁に使う人はコミュニケーション活動が活発でメッセンジャーも使うと考られる.

変量 Y のカテゴリー値 $b_1,\cdots,b_5$ を見ると変量 X と逆の傾向になっている．すなわち，メールを使わない人と，メールをとてもよく使う人のカテゴリー値が小さく，メールを使う頻度が中程度の人のカテゴリー値が相対的に大きい．これより，メールをとてもよく使う人はメッセンジャーを使うことが少なく，メールを使わない人はコミュニケーション活動が不活発なのでメッセンジャーも使わないと考えられる.

X と Y で逆の傾向があることには注目したい.

次に変量 U と V のカテゴリー値から，性別では男の方が，学部ではネットワーク情報学部の方がメッセンジャーの使用頻度が高いということがわかる.

4 種類の変量の中では，カテゴリー値のレンジ（最大と最小の差）が最も大きいのが学部であり，ネットワーク情報学部の特徴が現れていると見ることができる.

相関係数について メッセンジャーの使用度 Z に対する説明力を比較するのに，単相関係数・偏相関係数・レンジの 3 つの指標を考えることができるが，いずれも大きい順に，学部 V，性別 U，メールの使用度 Y，電話の使用度 X であった．すなわち，学部の違いが最も決定力が強いことになり，直感が裏付けられたと言えよう.

しかし，偏相関係数は最大値でも 0.44 とあまり大きくなく，この解析結果にそれほど説得力がないことを意味している．また，重相関係数は 0.5 に近いが，相関が強いといえるほどの値ではない．このことは，変数 Z の変動の説明に変量 X, Y, U, V だけでは十分でないことを意味している．分析方法の最大の問題は，4.1 節の最後に述べたように，変量 Z の恣意的な“数量化”である.

今回の解析結果は華々しい成果こそなかったが，方法論の紹介には十分な役割を果たしたと言えよう.

あとがき

今まで高度な分析はやったことがなく，数学が苦手だったこともあり勉強にはとても時間がかかっていた．理解もしなければならないし，報告会のために分析に取り掛からなければならないし，と板ばさみになりつつ，多くの人の協力で何とかここまでやってこれた．この分析を行ったことで，カテゴリカルな要素でもきちんと計算することで分析や予測が可能となっていく過程が大変面白く感じ，数学にも面白味を感じるようになれた．夜遅くまで計算し相談しあった経験は，大学生活の中でも屈指の思い出になるであろう．（永添）

3 年次のプロジェクトの授業でアンケート調査を行い，はじめて数量化という分析方法を知ることができた．アンケートの集計を行うだけでも大変な作業だったが，そこから数量化 I 類の分析をしていく過程で，計算が合わなかったり思わしい結果が出なかったりなど苦労の連続だった．しかし，苦労をしながらもこの分析を続けていくうちに，数量化についてさまざまな知識を得ることができた．このことは，楽しくもありとても有益なものだった．今回行った数量化という手法は色々なことに使えると思うので，この経験を生かしていきたい.（村上）

この手法は統計数理研究所の所長であった林 知己夫先生によって開発されたものである．昭和 40 年頃にこの手法が朝日新聞社の広告注目率調査に応用されたとき，混合型プログラムを作成することで協力した経験がある（学部 3–4 年生のとき）．その後，本学に赴任して，入試採点における選択科目の得点補正にこの手法を採用し，不公平是正に貢献できたことは幸いであった．このたび久しぶりにこの手法と再会し，学生諸君に手ほどきすることができて楽しかった.（佐藤）

末筆ながら，アンケート調査のために授業時間を割いていただいた高柳美香，石鎚英也両先生に感謝する.

参考文献

[1] 柳井晴夫，岩坪秀一,「複雑さに挑む科学–多変量解析入門」，ブルーバックス no.297，講談社，1976.
[2] 田中 豊，脇本和昌,「多変量統計解析法」，現代科学社，1983.
[3] 栗原考次,「データの科学」，放送大学印刷教材，放送大学教育振興会，2001.