数量化 2 類における追加情報検定と変数選択法

菅　民郎
中央大学理工学研究科 後期博士課程 数学専攻

1　序論

数量化法（quantification methods）には, 1 類, 2 類, 3 類, 4 類等複数の方法があり, アンケート調査などの質的データの分析に欠くことのできない解析手法である. ダミー変数の導入による質的データの数値化 { 1, 0 } により, 回帰分析を行うのが数量化 1 類, 判別分析を行うのが数量化 2 類と理解できる.

本論文では, 2 群の場合における数量化 2 類における追加情報の検定に対して, サンプルスコアに基づく統計量を提案する. この検定統計量が回帰分析や判別分析で知られている追加情報の検定統計量と同等の関係にあることを示す. この結果, 検定統計量の F 近似の妥当性は, 残差項の正規近似の度合いによって検証でき, これによってそのときの検定統計量の F 近似のよさを数値的に検証する. また, 回帰分析における追加情報検定統計量を利用した逐次変数選択法や, モデル選択基準に基づく変数選択法を 2 類に反映させ, 数値的検証を試みる. これらの結果は菅 (2009) として, 応用統計学会誌に掲載される. また多群の場合への拡張についても, 検討を進めている.

2　数量化 2 類とサンプルスコア

各個体について, 目的変数と q 個の説明変数（要因アイテムと呼ぶことにする）が観測されているとする. ただし, 目的変数は 2 値の群データ, q 個の要因アイテムはそれぞれ c_j $(j = 1, 2, \ldots, q)$ 個の選択肢を持つ質的データである. ここでデータを次の記号を用いて表す. 各群の個体数を n_1, n_2, 総個体数を n とする. g 番目群の i 番目の個体における目的変数の観察データを $y_i^{(g)}$ とする. また, j 番目アイテムの k 番目カテゴリーに対する反応について, 次のようなダミー変数を定義する.

$$x_{ijk}^{(g)} = \begin{cases} 1, & g \text{ 群の } i \text{ 番目の個体がアイテム } j \text{ のカテゴリー } k \text{ に反応するとき,} \\ 0, & \text{その他のとき.} \end{cases}$$

ここで, $g = 1, 2,\ \ i = 1, 2, \ldots, n_g,\ \ j = 1, 2, \ldots, q,\ \ k = 1, 2, \ldots, c_j$, である.

各個体が各々のアイテムのどのカテゴリーに反応したかを知ったとき, その情報にもとづいて目的変数を予測したい. そのため, 数量化 2 類におけるダミー変数の線形モデル

$$y_i^{(g)} = \sum_{j=1}^{q} \sum_{k=1}^{c_j} a_{jk} \left(x_{ijk}^{(g)} - \bar{x}_{jk} \right) + \varepsilon_i^{(g)} \tag{1}$$

を考える. ただし a_{jk} は線形式における係数, $\bar{x}_{jk}$ はダミー変数の平均, $\varepsilon_i^{(g)}$ は誤差である. どのアイテムにおいても, ダミー変数間の k に関する和について $\sum c_j x_{ijk}^{(g)} = 1$ の関係がある. この関係のもとで, 数量化 2 類における係数 a_{jk} は一般性を失うことなく, 各アイテムの任意のカテゴリーを除外したダミー変数で考えてよく, 以下では第一カテゴリーを除外した場合の a_{jk} の解法を検討する. なお, 除外したダミー変数の係数 a_{j1} は 0 とする.

数量化 2 類については, 林 (1993), 田中 (1983), 岩坪 (1987) などにおいて明記されている. ここではサンプルスコアについてまとめておく. 数量化 2 類において求められる合成変量 $\hat{y}_i^{(g)}$ を $\hat{y}_i^{(g)} = \sum_{j=1}^{q} \sum_{k=2}^{c_j} a_{jk} \{ x_{ijk}^{(g)} - \bar{x}_{jk} \}$ と表し, サンプルスコアと呼ぶことにする. (1) 式の係数ベクトル$\boldsymbol{a}$は, サンプルスコアと群との関係を示す相関比が最大になるように定められる. るサンプルスコア $\hat{y}_i^{(g)}$ の全体変動及び群間変動をそれぞれ $s_{\hat{y}}^2$ および s_b^2 とする. これらは次式で示せる.

$$s_{\hat{y}}^2 = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\hat{y}_i^{(g)} - \bar{\hat{y}})^2 = \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} = \boldsymbol{a}'T\boldsymbol{a}, \quad s_b^2 = \sum_{g=1}^{G} n_g(\bar{\hat{y}}^{(g)} - \bar{\hat{y}})^2 = \boldsymbol{a}'B\boldsymbol{a}. \tag{2}$$

ここに, T, B は全体変動行列, 群間変動行列であって, $T = \tilde{X}\tilde{X}'$, $B = \{n_1 n_2/n\}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'$ である. このとき相関比 η^2 は $\eta^2 = s_b^2/s_{\hat{y}}^2 = (\boldsymbol{a}'B\boldsymbol{a})/(\boldsymbol{a}'T\boldsymbol{a})$ で定義される. 最大となる相関比 η^2 は

$$\eta^2 = \frac{n_1 n_2}{n}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'T^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) \tag{3}$$

で与えられる.

η^2 が最大となる係数ベクトルは, サンプルスコアの平均が 0, 標準偏差が 1, すなわちサンプルスコアの偏差平方和が n となるように定めることにする. このような係数ベクトル $\boldsymbol{a}$ は

$$\boldsymbol{a} = \sqrt{n}\,\{(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'T^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})\}^{-1/2} \cdot T^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) \tag{4}$$

で与えられる.

3　数量化 2 類と回帰分析の関係について

数量化 2 類を実施したデータに回帰分析を適用したとき, 両者にどのような関係があるかを調べる. このとき, 回帰分析における目的変数に対して, どのような得点化をすべきかが問題になる. 判別に関係した得点化であるので, 群のみに依存した得点化

$$\boldsymbol{y} = (y_1^{(1)}, \ldots, y_n^{(1)}, y_1^{(2)} \ldots, y_n^{(2)})' = \begin{pmatrix} d_1\,\mathbf{1}_{\mathbf{n_1}} \\ d_2\,\mathbf{1}_{\mathbf{n_2}} \end{pmatrix} \tag{5}$$

を考える. ここに, $\mathbf{1_n}$ は要素がすべて 1 の n 次元ベクトルである. ここでは $\boldsymbol{y}$ について中心化したデータ $\tilde{\boldsymbol{y}}$ を

$$\tilde{\boldsymbol{y}} = (y_1^{(1)}, \ldots, y_n^{(1)}, y_1^{(2)} \ldots, y_n^{(2)})' = \begin{pmatrix} \sqrt{\dfrac{n_2}{n_1}}\,\mathbf{1}_{\mathbf{n_1}} \\ -\sqrt{\dfrac{n_1}{n_2}}\,\mathbf{1}_{\mathbf{n_2}} \end{pmatrix} \tag{6}$$

と定める.

このとき, 次に示す 2 点が成立することはよく知られている.

< 1 >　(1) 式のカテゴリースコアベクトルが回帰係数ベクトルと比例している（例えば, Anderson (2003)).

< 2 >　数量化 2 類における相関比を η^2, 回帰における決定係数 (重相関係数の二乗) を R^2 とすると, 両者は等しい. すなわち, $\eta^2 = R^2$ である.

これらのことは, より具体的には, 次のように述べられる. 数量化 2 類を実施したデータにおける回帰モデルを

$$y_i^{(g)} = b_0 + \sum_{j=1}^{q}\sum_{k=2}^{c_j} b_{jk}x_{ijk}^{(g)} + \varepsilon_i^{(g)} \tag{7}$$

とする. b は回帰係数, b_0 は定数項である. 目的変数の観察データ $y_i^{(g)}$ から平均 $\bar{y}$ を引いた偏差ベクトルを $\tilde{\boldsymbol{y}}$, 係数ベクトルを $\boldsymbol{b}$, p 次元ダミー変数における偏差データ行列を $\tilde{X}$ とする. このとき, 係数ベクトルの最小 2 乗推定量は

$$\boldsymbol{b} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\boldsymbol{y}} = \frac{n_1 n_2}{n}T^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) \tag{8}$$

で表せる. (4) 式の $\boldsymbol{a}$ と (8) 式の $\boldsymbol{b}$ は比例している.

回帰モデルの決定係数 (重相関係数の二乗)　R^2 は次式によって求められる.

$$R^2 = \frac{(y_i^{(g)}\text{と}\hat{y}_i^{(g)}\text{の偏差積和})^2}{(y_i^{(g)}\text{の偏差平方和}) \cdot (\hat{y}_i^{(g)}\text{の偏差平方和})} = \frac{\{\displaystyle\sum_{g=1}^{2}\sum_{i=1}^{n_g}(y_i^{(g)} - \bar{y})(\hat{y}_i^{(g)} - \bar{\hat{y}})\}^2}{\displaystyle\sum_{g=1}^{2}\sum_{i=1}^{n_g}(y_i^{(g)} - \bar{y})^2 \sum_{g=1}^{2}\sum_{i=1}^{n_g}(\hat{y}_i^{(g)} - \bar{\hat{y}})^2}. \tag{9}$$

これを計算すると

$$R^2 = \frac{n_1 n_2}{n}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'T^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) \tag{10}$$

となり, (3) 式の η^2 に一致する.

4　要因アイテムの目的変数への寄与度を調べる追加情報の検定

追加情報の検定とは, 説明変数 p 個の中の任意の変数の組 J を除いたモデル M と, モデル M に J を含めたモデル Ω との関係から導かれる検定統計量に基づき, 追加変数 J が目的変数へ有意に寄与しているかを調べる検定方法である.

数量化 2 類より導かれた (1) 式をモデル Ω と名付ける. モデル Ω においてアイテム j^* を除いた線形モデルをモデル M と名づける. モデル Ω の説明変数のアイテム数を q, ダミー変数の個数を p, モデル M のアイテム数を $q-1$, ダミー変数の個数を p', p 個の変数を用いたときのサンプルスコアの群内変動を $s^2_{w(\Omega)}$, p' 個の変数を用いたときのサンプルスコア群内変動を $s^2_{w(M)}$ とする.

このとき, モデル Ω のあるアイテム j^* が目的変数に有意に寄与しているかを調べる検定統計量として次式を提案する.

$$F_w = \frac{(s^2_{w(M)} - s^2_{w(\Omega)})/(p-p')}{s^2_{w(\Omega)}/(n-p-1)}. \tag{11}$$

回帰分析において説明変数を p' から p へ増加したとき, 加えた説明変数が目的変数に有意に寄与しているかどうかは, 次に示す検定統計量

$$F = \frac{(s^2_{e(M)} - s^2_{e(\Omega)})/(p-p')}{s^2_{e(\Omega)}/(n-p-1)} \tag{12}$$

の値より把握できる. F 値は追加した説明変数の係数が 0 であるという帰無仮説, 誤差 $\varepsilon_i^{(g)} \sim N(0, \sigma^2)$ のもとで, 自由度 $(p-p', n-p-1)$ の F 分布にしたがうことが知られている. ただし $s^2_{e(M)}, s^2_{e(\Omega)}$ は説明変数が p' 個のモデル M, p 個のモデル Ω に回帰分析を適用したときの残差平方和である.

(11) 式内の s^2_w と (12) 式内の s^2_e は等しいことを示せる (本論文, 定理 3.1 参照) ので, 2 類と回帰分析の追加情報の検定統計量が等しいことを指摘できる.

判別分析の場合, p' 個の変数に $p-p'$ 個の変数を追加したときの有意性検定として

$$F_D = \frac{\frac{n_1 n_2}{n}\{D^2_p - D^2_{p'}\}}{n-2+\frac{n_1 n_2}{n}D^2_{p'}} \cdot \frac{n-p-1}{p-p'} \tag{13}$$

が自由度 $p-p', n-p-1$ の F 分布に従うことを用いる方法が知られている. ここに, D^2_p は p 個の変数を用いたときのマハラノビスの距離の二乗で, $D^2_{p'}$ は p' 個の変数を用いたときのマハラノビスの二乗である. 判別分析における F_D が回帰における F に対応するものであるが, 2 類の F_w が F_D に一致していることを指摘できる (本論文, 定理 3.2 参照).

このことより, 回帰分析、判別分析同様に数量化 2 類でも F_w を用いて, 追加したアイテム j^* の有意性を調べることができると考えられる. 2 類における検定統計量 F_w の F 近似の妥当性は, F_w と F との同等性, F_W と F_D との同等性, 残差項の正規近似の妥当性に帰着できると判断する. 本論文では F_w の F 近似のよさを数値実験で検証した. 残差項正規近似の数値的検証は本論文の 6 節, F_w の F 近似数値実験は本論文の 7 節で示す.

具体的な検定手順は, 追加したアイテム j^* にある全てのダミー変数の係数が 0 であるという帰無仮説, 誤差 $\varepsilon_i^{(g)} \sim N(0, \sigma^2)$ のもとで, 有意水準 α, 自由度 $(c_{j^*}-1, n-p-1)$ の F 分布の限界値 F_0 と F_w 値を比較し, $F_w \geqq F_0$ ならば帰無仮説を棄却する. すなわちアイテム j^* は目的変数に有意に寄与していると判断する. ただし, c_{j^*} は追加アイテム j_* のカテゴリー数, p は追加アイテムを含んだ全てのダミー変数の総個数である.

5　数量化 2 類におけるモデル選択基準と変数選択

数量化 2 類において, 要因アイテムの候補の中から最良なものを選択してモデル式を求める方法について考える. これまで, 2 類の数量化法と回帰との間には密接な関連があることを見てきたが, これらの関連を通して,

回帰分析などで用いられている *AIC*(*Akaike*(1978)) や C_p(*Mallows*(1973)) などのモデル選択基準を 2 類に反映させることが考えられる.

このような考えは, 例えば芳賀敏郎 (1990) においても考えられている. しかし, このような形式的な展開の妥当性については研究されていない. ここでは, 回帰によるアプローチにおいて, 残差の変動が正規に近いものであれば, *AIC* や C_p などのモデル選択基準が適用でき, 数値実験を通してこれらの基準が有用であることを指摘できる.

数量化 2 類において, 要因アイテムの候補の中から最良なアイテムを見出す選択方法について考える. 回帰分析などで利用されている逐次選択方法は (11) 式の F を用いて変数の取り込みや掃き出し行うが, 2 類の F_w と回帰の F が等しいことから, 2 類も F_W を適用して逐次変数選択法が行なえることを指摘できる.

6 多群への拡張

多群の場合への拡張においては, 多変量回帰モデルや多群の場合の判別分析との関係を調べることになる. 多変量回帰との関連においては, 目的変数の得点化が基本になる. 多群判別の係数ベクトルと多変量回帰の係数ベクトルが同じになるための得点化については, *Hstieetal.*(1994) によって与えられている. この得点化を利用して, 2 群の場合におけるような追加情報検定統計量の一致性, 検定統計量の近似分布の検証, 変数選択法などについて, 拡張結果を得ている. これらの結果は藤越康祝氏との共同研究である.

7 結論

本論文では, 主として 2 群の場合の数量化 2 類の統計的推測に関連して, 数量化 2 類と回帰との間の密接な関連を調べた. また, 追加情報の検定について, 適当な標本数があれば, 検定統計量が F 分布で近似できることを数値的に検証した. さらに, 回帰分析におけると同様の変数選択法が適用できることを指摘した. 回帰との関連では, 目的変数の得点化（数量化）も基本になるが, 多群の場合, このことがより重要になる. この結果を利用して多群の数量化 2 類への拡張を進めているが, 今後さらに発展させたい.

参考文献

[1] Akaike, H. (1973). Informaiton theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, (B. N. Petrov and F.Csáki,eds.), 267–81, Budapest: Akadémia Kiado.

[2] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). John Wiley & Sons, New York.

[3] Hastie, T. Tibshirani, R. and Buja, A (1994). Flexible discriminant analysis by optional scoring . J. Amer. Sfatiot. Assoc., 89, 1225-1270.

[4] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661-675.

[5] Rencher, A. C. (2002). Methods of Multivariate Analysis. Wiley Series Probability and Statistics, Wiley-lnterscience.

[6] 芳賀敏郎 (1990). 数量化 1 類と数量化 2 類におけるアイテムの選択. 人間行動計量分析 (柳井・岩坪・石塚編),155-171.

[7] 林知己夫 (1993). 数量化-理論と方法. 朝倉書店.

[8] 岩坪秀一 (1987). 統計ライブラリー数量化法の基礎. 朝倉書店.

[9] 菅民郎 (2009). 数量化 2 類における追加情報検定と変数選択法. 応用統計学誌第 38 巻に掲載予定.

[10] 田中豊・脇本和昌 (1983). 多変量統計解析法. 現代数学社.