

数量化 1 類

従属変数（外的基準値という）が連続量で与えられていて、独立変数が質的データである場合において、外的基準値を予測するのに最適な質的データの数量化を考える。 i 番目の個体の外的基準値を $y(i)$ で、アイテム j におけるカテゴリ k の値を $n_i(jk)$ で表す¹。ここで、 $n_i(jk)$ の値は次のように定める。

$$n_i(jk) = \begin{cases} 1 & i \text{ 番目の個体がアイテム } j \text{ のカテゴリ } k \text{ を選択しているとき} \\ 0 & i \text{ 番目の個体がアイテム } j \text{ のカテゴリ } k \text{ を選択していないとき} \end{cases}$$

それぞれのアイテムから 1 つのカテゴリが選択されるものとする。個体の総数を m 、アイテムの総数を n 、アイテム j のカテゴリの数を N_j とおく。

$n_i(jk)$ は次式を満たす。

$$\sum_{k=1}^{N_j} n_i(jk) = 1$$

いま、アイテム j におけるカテゴリ k の値に適当な数値 $x(jk)$ を割り当てて外的基準値 $y(i)$ を次式で推定することを考える。

$$y(i) \approx \hat{y}(i) = c + \sum_{j=1}^n \sum_{k=1}^{N_j} x(jk) \cdot n_i(jk) \quad (1)$$

ここで、 c は適当な定数である。

(1) 式による推定の誤差を次式で表す。

$$Q = \sum_{i=1}^m (y(i) - \hat{y}(i))^2 \quad (2)$$

(2) 式の値を最小にする c は次式

$$c = \bar{y} - \sum_{j=1}^n \sum_{k=1}^{N_j} x(jk) \cdot \bar{n}(jk) \quad (3)$$

で与えられ、 $x(jk)$ は次式を満たすものとして与えられる。

$$\mathbf{Ax} = \mathbf{b} \quad (4)$$

¹ ここでの記法は岩坪 (1987) に従うものである。

ここで、

$$\bar{y} = \frac{1}{m} \sum_{i=1}^m y(i)$$

$$\bar{n}(jk) = \frac{1}{m} \sum_{i=1}^m n_i(jk)$$

である。

行列

$$\mathbf{A} = (a(jk, j'k'))$$

およびベクトル

$$\mathbf{b} = (b(jk))$$

の要素は次式で与えられる。

$$a(jk, j'k') = \frac{1}{m} \sum_{i=1}^m \{n_i(jk) - \bar{n}(jk)\} \{n_i(j'k') - \bar{n}(j'k')\}$$

$$b(jk) = \frac{1}{m} \sum_{i=1}^m \{y(i) - \bar{y}\} \{n_i(jk) - \bar{n}(jk)\}$$

(4) 式の解を、岩坪 (1987) はムーア・ペンローズ逆行列 \mathbf{A}^+ を用いて

$$\mathbf{x} = \mathbf{A}^+ \mathbf{b}$$

で与え、議論を展開している。

(4) 式の 1 つの解において、同じアイテム内のカテゴリについて定数を加えたものも (4) 式の解である (証明については付録を参照)。各アイテムごとの $x(jk)$ の平均値が 0 になるように定数を加える (標準化)。アイテム j における平均値は次式で与えられる。

$$\frac{1}{m} \sum_i \left\{ \sum_k n_i(jk) x(jk) \right\} = \sum_k \left[\left\{ \frac{1}{m} \sum_i n_i(jk) \right\} x(jk) \right]$$

$$= \sum_k \bar{n}(jk) x(jk)$$

上式の値を $x(jk)$ から引いた値

$$x^*(jk) = x(jk) - \sum_k \bar{n}(jk) x(jk)$$

は平均値が 0 になる。すなわち次式が成り立つ。

$$\frac{1}{m} \sum_i \left\{ \sum_k n_i(jk) x^*(jk) \right\}$$

$$= \frac{1}{m} \sum_i \left[\sum_k n_i(jk) \left\{ x(jk) - \sum_k \bar{n}(jk) x(jk) \right\} \right]$$

$$\begin{aligned}
 &= \frac{1}{m} \sum_i \left\{ \sum_k n_i(jk) x(jk) \right\} - \frac{1}{m} \sum_i \left[\left\{ \sum_k \bar{n}(jk) x(jk) \right\} \cdot \sum_k n_i(jk) \right] \\
 &= \sum_k \bar{n}(jk) x(jk) - \frac{1}{m} \sum_i \left[\left\{ \sum_k \bar{n}(jk) x(jk) \right\} \times 1 \right] \\
 &= \sum_k \bar{n}(jk) x(jk) - \sum_k \bar{n}(jk) x(jk) \\
 &= 0
 \end{aligned}$$

標準化された解 $\mathbf{x}^* = (x^*(jk))$ による外的基準値 $y(i)$ の推定値 $\hat{y}(i)$ は次式で与えられる。

$$\hat{y}(i) = \bar{y} + \sum_{j=1}^n \sum_{k=1}^{N_j} x^*(jk) \cdot n_i(jk)$$

これは、(3) 式より以下のように導くことができる。

$$\begin{aligned}
 c &= \bar{y} + \sum_{j=1}^n \sum_{k=1}^{N_j} x^*(jk) \cdot \bar{n}(jk) \\
 &= \bar{y} + \sum_{j=1}^n \sum_{k=1}^{N_j} x^*(jk) \cdot \frac{1}{m} \sum_{i=1}^m n_i(jk) \\
 &= \bar{y} + \sum_{j=1}^n \left\{ \frac{1}{m} \sum_{i=1}^m \sum_{k=1}^{N_j} n_i(jk) x^*(jk) \right\} \\
 &= \bar{y}
 \end{aligned}$$

付 録

$\mathbf{x} = (x(jk))$ が (4) 式の解であるとき、アイテム j' のカテゴリ k' に $c_{j'}$ を加えたものを

$$\mathbf{x}_1 = (x_1(jk))$$

とおく。

$$x_1(jk) = \begin{cases} x(jk) + c_{j'} & j = j' \text{ のとき} \\ x(jk) & j \neq j' \text{ のとき} \end{cases}$$

である。このとき

$$\mathbf{x}_1 = \mathbf{x} + \mathbf{d}$$

とおけば、

$$\mathbf{d}' = \begin{bmatrix} 0 & \cdots & 0 & \underbrace{c_{j'} \cdots c_{j'}}_{\text{項目 } j' \text{ のカテゴリ}} & 0 & \cdots & 0 \end{bmatrix}$$

となっている。よって

$$\mathbf{Ax}_1 = \mathbf{Ax} + \mathbf{Ad}$$

となる。 \mathbf{Ad} の (jk) 行目の値は次式で与えられる。

$$\begin{aligned} \sum_{k'=1}^{N_{j'}} a(jk, j'k') \cdot c_{j'} &= c_{j'} \cdot \sum_{k'=1}^{N_{j'}} \left[\frac{1}{m} \sum_i \{n_i(jk) - \bar{n}(jk)\} \{n_i(j'k') - \bar{n}(j'k')\} \right] \\ &= \frac{c_{j'}}{m} \sum_i \left[\{n_i(jk) - \bar{n}(jk)\} \cdot \sum_{k'} \{n_i(j'k') - \bar{n}(j'k')\} \right] \end{aligned}$$

ここで、

$$\begin{aligned} \sum_{k'} \{n_i(j'k') - \bar{n}(j'k')\} &= \sum_{k'} n_i(j'k') - \sum_{k'} \bar{n}(j'k') \\ &= 1 - \sum_{k'} \left\{ \frac{1}{m} \sum_i n_i(j'k') \right\} \\ &= 1 - \frac{1}{m} \sum_i \left\{ \sum_{k'} n_i(j'k') \right\} \\ &= 1 - \frac{1}{m} \sum_{i=1}^m 1 \\ &= 1 - \frac{1}{m} \times m \\ &= 0 \end{aligned}$$

すなわち、

$$\mathbf{Ad} = \mathbf{0}$$

となる。よって、

$$\mathbf{Ax}_1 = \mathbf{Ax} = \mathbf{b}$$

の成り立つことが分る。

参考文献

岩坪秀一 1987 「数量化法の基礎」 朝倉書店