2 値データの相関係数のベイズ推定

○岡本安晴

(日本女子大学心理学科)

キーワード: 四分相関係数、MCMC、事後分布

Bayesian estimation of correlation for binary data

Yasuharu OKAMOTO

(Dept. Psychology, Japan Women's Univ.)

Key Words: Tetrachoric correlation, MCMC, posterior distribution

目 的

2次元正規分布に従う変量 (U, V) のデータが2値データ (X, Y) として得られたとき、 X と Y の相関係数(ファイ係数 ϕ)は U と V の真の相関係数 ρ より小さくなるが、条件によっては ϕ は ρ より著しく小さくなる(岡本; 1987, 1988)。2値データから ρ を推定する値として四分相関係数(tetrachoric correlation coefficient)があり、最尤法(ML)が用いられる。Choiら(2011)は、MLより優れた方法としてベイズの方法を提案した。Choiらの方法は、2値データを含む一般に多値カテゴリーデータに対するもので、この場合の推定値は多分相関係数(polychoric correlation coefficient)と呼ばれる。Choiらの方法は、MLの2段推定法において、1段目のカテゴリー境界の推定は2段ML法の場合と同じく度数分布から累積正規分布関数の逆関数の値として求めているが、2段目の ρ の推定を最尤法ではなくベイズ的に行うものである。しかし、この方法は、事後確率分布においてカテゴリー境界を固定した条件付き事後確率分布によって分析を行うもので、事後分布全体を考慮したベイズ的分析ではない。Choiらのように2段に分けて後段のみベイズ的に行うのではなく、カテゴリー境界と相関係数の両方を同時に扱うベイズの方法を提案する。

モデル

変量 U および V のカテゴリー境界を

$$-\infty = a_0 < a_1 < \dots < a_{K-1} < a_K = +\infty$$

および

$$-\infty = b_0 < b_1 < \dots < b_{H-1} < b_H = +\infty$$

とおく。このとき、

$$X = k, \quad a_{k-1} < U < a_k \text{ のとき}$$

$$Y = h, \quad b_{h-1} < V < b_h \text{ のとき}$$

であるとする。 U と V の同時分布を

$$(U, V) \sim N_2(\mathbf{0}, \Sigma), \quad \mathbf{0} = (0, 0), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

とおいても一般性は失われない。 $N_2(\mathbf{0}, \Sigma)$ の確率密度関数を $f(u, v | r)$ とおけば、

$$P(X = k, Y = h | \rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1})$$

$$= \int_{a_{k-1}}^{a_k} \int_{b_{h-1}}^{b_h} f(u, v | \rho) du dv$$

である。 $X = k$ かつ $Y = h$ であった度数を N_{kh} とおくと

$$P(N_{11}, \dots, N_{KH} | \rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1})$$

$$\propto \prod_{k=1}^K \prod_{h=1}^H P(X = k, Y = h | \rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1})^{N_{kh}}$$

である。

事前分布を

$$p_0(\rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1})$$

とおき、尤度を

$$L(\rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1} | N_{11}, \dots, N_{KH}) \\ = \prod_{k=1}^K \prod_{h=1}^H P(X = k, Y = h | \rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1})^{N_{kh}}$$

おけば、事後分布は

$$P(\rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1} | N_{11}, \dots, N_{KH}) \\ \propto p_0(\rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1}) \\ \times L(\rho, a_1, \dots, a_{K-1}, b_1, \dots, b_{H-1} | N_{11}, \dots, N_{KH})$$

と表される。

事後分布をMCMCの一種であるMetropolis-within-Gibbsアルゴリズムによって推定する。提案分布として正規分布を用いるため、パラメータ値の制約条件を事前分布によって設定する。以下の更新ステップを $t = 0$ から $t = N - 1$ まで繰り返すことによって N ステップ分の事後分布のサンプル値を得る。

0. $t \leftarrow 0$ とおく。

1. サンプルの候補値 $y \sim N(\rho^{(t)}, \sigma_r^2)$ とそのaccept確率 α_r を求める。確率 α_r で $\rho^{(t+1)} = y$ 、確率 $1 - \alpha_r$ で $\rho^{(t+1)} = \rho^{(t)}$ とする。

2. $k = 1$ とおき、 $k = 1$ から K までステップ2a~2bを繰り返す。

2a. サンプルの候補値 $y \sim N(a_k^{(t)}, \sigma_{ak}^2)$ とそのaccept確率 α_{ak} を求める。確率 α_{ak} で $a_k^{(t+1)} = y$ 、確率 $1 - \alpha_{ak}$ で $a_k^{(t+1)} = a_k^{(t)}$ とおく。

2b. $k \leftarrow k + 1$ とおく。 $k < K$ ならば2aに戻る。

3. $h = 1$ とおき、 $h = 1$ から H までステップ3a~3bを繰り返す。

3a. サンプルの候補値 $y \sim N(b_h^{(t)}, \sigma_{bh}^2)$ とそのaccept確率 α_{bh} を求める。確率 α_{bh} で $b_h^{(t+1)} = y$ 、確率 $1 - \alpha_{bh}$ で $b_h^{(t+1)} = b_h^{(t)}$ とおく。

3b. $h \leftarrow h + 1$ とおく。 $h < H$ ならば3aに戻る。

4. $t \leftarrow t + 1$ とおく。 $t < N$ ならばステップ1に戻る。

結果と考察

$K = H = 2$ の場合の分析例を発表会場で示すが、相関係数 ρ およびカテゴリー境界 a_1 と b_1 の事後分布は満足のいくものであり、提案したベイズの方法の有用性を示すものである。

引用文献

Choi, J., et al. (2011). *Jour. Educ. and Beh. Stat.*, 36, 523-549.

岡本安晴(1987). (<http://ci.nii.ac.jp/naid/110000976183>)

岡本安晴(1988). (<http://ci.nii.ac.jp/naid/110000976190>)