大学院システム情報科学府 基礎共通科目　　2011 年 6 月 23 日

確率・統計特論 第 10 回

来嶋 秀治 (Shuji Kijima)

締切: 2011 年 6 月 29 日講義終了時回収
注意: 参照した文献等の情報を必ず記載すること.

今日の話題: 線形回帰, 多次元分布, AIC

1. 回帰分析

1-i. 線形回帰分析: 単回帰分析. 確率変数 X, Y は未知パラメータ α, β に対して, $Y = \alpha X + \beta$ を満たすと想定する. いま, k 個の標本 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_k)$ を得たとする. **最小二乗推定量** (*least square estimator*) $(\widehat{\alpha}, \widehat{\beta})$ を

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha', \beta'} \sum_{i=1}^{k} (y_i - (\alpha' x_i + \beta_i'))^2$$

を満たすものとする. これを解析的に解くと, 具体的には

$$\widehat{\alpha} = s_{x,y}/s_x^2, \quad \widehat{\beta} = \overline{y} - \widehat{\beta}\overline{x}$$

となる. ただし

$$\overline{x} = \sum_{i=1}^{n} x_i/n, \quad \overline{y} = \sum_{i=1}^{n} y_i/n, \quad \overline{x^2} = \sum_{i=1}^{n} x_i^2/n, \quad \overline{xy} = \sum_{i=1}^{n} x_i y_i/n, \quad s_x^2 = \overline{x^2} - \overline{x}^2 \quad s_{x,y} = \overline{xy} - \overline{xy}$$

とする. remember $\mathrm{Cov}[XY] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$.

1-ii. 線形回帰分析: 重回帰分析. 確率変数 $X_1, \ldots, X_n$ は未知パラメータ $\alpha_1, \ldots, \alpha_n, \beta$ に対して, $Y = \sum_{i=1}^{n} \alpha_i X_i + \beta$ を満たすと想定する. いま, k 個の標本 $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^k$ を得たとする. このとき, $\alpha_1, \ldots, \alpha_n, \beta$ を推定する手法が重回帰分析である ($\Longrightarrow$ 演習 1).

1-iii. その他の回帰分析. 一般に, 想定する関係式が線形である必要はない. たとえば, 2 つの確率変数 X, Y が未知パラメータ α, C に対して, $Y = CX^{\alpha}$ を想定した方がよい場合もある ($\Longrightarrow$ 演習 2).
線形回帰分析以外の回帰としてロジスティック回帰などが挙げられる.

2. 多次元分布

(i) 多次元正規分布 (multivariate normal distribution) $N(\boldsymbol{\mu}, \Sigma)$
($\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\top} \in \mathbb{R}^n$, $\Sigma = (\sigma_{i,j}) \in \mathbb{R}^{n \times n}$: 分散共分散行列 (variance-covariance matrix).)

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\top}\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \qquad (\boldsymbol{x} \in \mathbb{R}^n)$$

ただし

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n}^2 & \cdots & \sigma_n^2 \end{pmatrix}$$

(ii) Dirichlet 分布 (Dirichlet distribution) $\mathrm{Dir}(\boldsymbol{\alpha})$ ($\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n_{\geq 0}$)
標本空間 $\Omega = \{\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n_{>0} \mid x_1 + \cdots + x_n = 1\}$,

$$f(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \qquad (\boldsymbol{x} \in \Omega)$$

ただし,

$$B(\boldsymbol{\alpha}) \;=\; \int_{\Omega} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \mathrm{d}\boldsymbol{x} \;=\; \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{n} \alpha_i)}$$

(iii) 多項分布 (multinomial distribution) $\mathrm{M}(\boldsymbol{p}, K)$ ($\boldsymbol{p} = (p_1, \ldots, p_n) \in \mathbb{R}^n_{>0}$, $\sum_{i=1}^n p_i = 1$)
標本空間 $\Omega = \{\boldsymbol{z} \in \mathbb{Z}^n_{\geq 0} \mid z_1 + \cdots + z_n = K\}$

$$f(\boldsymbol{z}) = \frac{K!}{z_1! z_2! \cdots z_n!} p_1^{z_1} \cdots p_n^{z_n} \qquad (\boldsymbol{z} \in \Omega)$$

※多次元正規分布, Dirichlet 分布は連続分布, 多項分布は離散分布であることに注意.

3. モデル選択 — オッカムの剃刀 (Occam's razor)

AIC (赤池情報量基準): 統計データ (k 個の標本)$\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$ に対して, モデル (確率分布)$f(\boldsymbol{x}; \theta)$ の AIC を

$$\begin{aligned} \mathrm{AIC} \;&:=\; -2 \sum_{i=1}^{n} \log L(\widehat{\theta}_1, \ldots, \widehat{\theta}_q; \boldsymbol{x}^i) + 2(q+2) \\ &=\; -2(\text{最大対数尤度}) + 2(\text{パラメータ数}) \end{aligned}$$

とする. AIC の小さなモデルが, より良いモデルと推定される.
※同様の概念として BIC (ベイズ情報量規準), MDL(最小記述長) などがある.

演習問題

演習 1. 正の確率変数 X と Y は未知パラメータ α と C に対して, $Y = CX^\alpha$ を満たすものと想定される. いま k 個の標本 $(x_1, y_1), \ldots, (x_k, y_k)$ が得られたとする. このとき α と C を推定せよ.
ヒント: $Y = CX^\alpha$ の両辺の対数を考えよ.

演習 2. 確率変数 X, Y, Z は未知パラメータ a, b, c に対して, $Z = aX + bY + c$ を満たすものと想定される. いま k 個の標本 $(x_1, y_1, z_1), \ldots, (x_k, y_k, z_k)$ が得られたとする. このとき a, b, c を推定せよ.

参考文献

樺島祥介, 北川源四郎, 甘利俊一, 赤池弘次, 下平英寿, 土谷隆 (編), 室田一雄 (編), 赤池情報量規準 AIC—モデリング・予測・知識発見, 共立出版 (2007).