

応用システム工学
第4回 指数型分布族と
一般化線形モデル
平成21年6月26日

統計モデル

- 線形モデルに対するデータの分析
 - 正規分布を持つ独立な確率変数 Y_i
 - $Y_i \sim N(\mu_i, \sigma^2)$
 - 期待値の線形モデル
 - $E(Y_i) = \mu_i = X_i^T \beta$
- 線形モデルの一般化
 - 正規分布以外の確率変数
 - 指数分布族(より広いクラスの分布)→本授業の対象
 - 確率変数と説明変数の関係が非線形
 - 連結関数の適用 $g(\mu_i) = X_i^T \beta$
 - 一般化加法モデル
 - 大量の計算処理必要→本授業では扱わない

説明変数: X
パラメータ: β

指数型分布族の確率変数

- 指数型分布族の条件
 - 確率変数 Y , 単一パラメータ θ の指数関数で表される確率分布を持つ→最尤推定に便利

$$f(y; \theta) = s(y)t(\theta)\exp[a(y)b(\theta)]$$

- 但し, a, b, s, t は既知の関数
- $s(y) = \exp[d(y)], t(\theta) = \exp[c(\theta)]$ とおく

$$\begin{aligned} f(y; \theta) &= \exp[d(y)]\exp[c(\theta)]\exp[a(y)b(\theta)] \\ &= \exp[a(y)b(\theta) + c(\theta) + d(y)] \end{aligned}$$

指数型分布族の確率変数

$$f(y; \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

- 正準形の分布: $a(y)=y$
- 分布の自然パラメータ: $b(\theta)$
- 局外パラメータ $\rightarrow \theta$ 以外のパラメータ
 - 既知として扱う必要あり(単一パラメータのため)
- 指数型分布族の例
 - 正規分布, ポアソン分布, ワイブル分布, 二項分布

指数型分布族としての正規分布

- 正規分布 $Y \sim N(\mu, \sigma^2)$
 - 対称な分布を持つ連続データに対するモデル
 - 確率密度関数

$$f(y; \mu) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]$$

$$= \exp\left[\log \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} - \frac{1}{2\sigma^2}(y - \mu)^2\right]$$

$$= \exp\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right]$$

次頁に
つづく

指数型分布族としての正規分布

- 確率密度関数

$$\begin{aligned} f(y; \mu) &= \exp \left[\frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right] \\ &= \exp[a(y)b(\mu) + c(\mu) + d(y)] \quad \Rightarrow a(y)b(\mu) = \frac{y\mu}{\sigma^2} \end{aligned}$$

- パラメータ μ

- 正準形 $a(y) = y$

- 自然パラメータ $b(\mu) = \frac{\mu}{\sigma^2}$

- 局外パラメータ σ^2

- その他の項 $c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2), d(y) = -\frac{y^2}{2\sigma^2}$

指数型分布族としてのポアソン分布

- ポアソン分布 $Y \sim \text{Poisson}(\theta)$
 - 計数データのモデルとして用いる
 - 確率変数: $Y \rightarrow$ 離散値, $y=1,2,3,\dots$ 等
 - 確率密度関数
$$f(y; \theta) = \frac{\theta^y e^{-\theta}}{y!}$$
$$= \exp(y \log \theta - \theta - \log y!)$$
$$= \exp[a(y)b(\theta) + c(\theta) + d(y)]$$
- 正準形: $a(y)=y$ である
- 自然パラメータ $b(\theta) = \log \theta$
- $c(\theta) = -\theta$, $d(y) = -\log y!$

指数型分布族としてのワイブル分布

- ワイブル分布
 - 故障時間(生存期間)の分布を表すモデル
 - 確率変数: y (時間)
 - パラメータ: λ (分布の形状), θ (尺度)
 - 確率密度関数

$$\begin{aligned} f(y; \mu, \theta) &= \frac{\lambda y^{\lambda-1}}{\theta^\lambda} \exp\left[-\left(\frac{y}{\theta}\right)^\lambda\right] \\ &= \exp\left[\log \lambda + (\lambda - 1)\log y - \lambda \log \theta - \left(\frac{y}{\theta}\right)^\lambda\right] \end{aligned}$$

指数型分布族としてのワイブル分布

$$\begin{aligned} f(y; \mu, \theta) &= \exp \left[\log \lambda + (\lambda - 1) \log y - \lambda \log \theta - \left(\frac{y}{\theta} \right)^\lambda \right] \\ &= \exp [a(y)b(\theta) + c(\theta) + d(y)] \end{aligned}$$

- パラメータ: θ
- $a(y) = y^\lambda$ $\lambda = 1$ のとき正準形となる
- 自然パラメータ $b(\theta) = -\theta^{-\lambda}$
 - 局外パラメータ: λ
- その他の項
$$c(\theta) = -\lambda \log \theta$$
$$d(y) = \log \lambda + (\lambda - 1) \log y$$

指数型分布族としての二項分布

- 二項分布 $Y \sim \text{Binominal}(n, \pi)$
 - 二値(0,1)をとるn回の独立試行に対する分布
 - 確率変数: $Y \rightarrow 1$ の回数。離散値, $y=1, 2, 3, \dots, n$
 - 二項係数 $\rightarrow$ n個からy個を選ぶ組み合わせの数

$$\binom{n}{y} = {}_n C_y = C(n, y) = \frac{n!}{y!(n-y)!}$$

- 確率密度関数

$$f(y; \pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}$$

指数型分布族としての二項分布

- 確率密度関数

$$\begin{aligned} f(y; \pi) &= \exp \left[\log \binom{n}{y} + y \log \pi + (n - y) \log(1 - \pi) \right] \\ &= \exp \left[y \log \pi - y \log(1 - \pi) + n \log(1 - \pi) + \log \binom{n}{y} \right] \\ &= \exp \left[y \log \frac{\pi}{1 - \pi} + n \log(1 - \pi) + \log \binom{n}{y} \right] \\ &= \exp [a(y)b(\pi) + c(\pi) + d(y)] \end{aligned}$$

指数型分布族としての二項分布

- パラメータ π
- 正準形 $a(y) = y$
- 自然パラメータ $b(\pi) = \log \frac{\pi}{1-\pi}$
 - 局外パラメータ n
- その他の項 $c(\pi) = n \log(1-\pi)$
$$d(y) = \log \binom{n}{y}$$

指数型分布族の性質

- 確率変数 $a(Y)$ の期待値を求める

$$E[a(Y)] = \int a(y) f(y; \theta) dy$$

– 確率密度関数の性質(全体の確率は1)

$$\int f(y; \theta) dy = 1$$

- パラメータに対する微分の性質

$$\frac{d}{d\theta} \int f(y; \theta) dy = \frac{d}{d\theta} 1 = 0$$

$$\Rightarrow \int \frac{d}{d\theta} f(y; \theta) dy = 0$$

指数型分布族の性質

- 指数型分布族の確率密度関数のパラメータに対する微分

$$f(y, \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

$$\begin{aligned}\frac{d}{d\theta} f(y, \theta) &= \frac{d}{d\theta} \exp[a(y)b(\theta) + c(\theta) + d(y)] \\ &= [a(y)b'(\theta) + c'(\theta)] \\ &\quad \times \exp[a(y)b(\theta) + c(\theta) + d(y)] \\ &= [a(y)b'(\theta) + c'(\theta)] f(y, \theta)\end{aligned}$$

指数型分布族の性質

$$\begin{aligned}\int \frac{d}{d\theta} f(y, \theta) dy &= \int [a(y)b'(\theta) + c'(\theta)] f(y, \theta) dy \\ &= b'(\theta) \int a(y) f(y, \theta) dy \\ &\quad + c'(\theta) \int f(y, \theta) dy \\ &= 0\end{aligned}$$

$$b'(\theta)E[a(Y)] + c'(\theta) = 0$$

$$E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}$$

指数型分布族の性質

－ 確率変数 $a(Y)$ の分散

$$\text{var}[a(Y)] = \int \{a(y) - E[a(y)]\}^2 f(y; \theta) dy$$

$$\begin{aligned} \frac{d^2}{d\theta^2} f(y, \theta) &= \frac{d^2}{d\theta^2} \exp[a(y)b(\theta) + c(\theta) + d(y)] \\ &= \frac{d}{d\theta} \{[a(y)b'(\theta) + c'(\theta)] f(y, \theta)\} \\ &= [a(y)b''(\theta) + c''(\theta)] f(y, \theta) + [a(y)b'(\theta) + c'(\theta)] \frac{d}{d\theta} f(y, \theta) \\ &= [a(y)b''(\theta) + c''(\theta)] f(y, \theta) + [a(y)b'(\theta) + c'(\theta)] [a(y)b'(\theta) + c'(\theta)] f(y, \theta) \\ &= [a(y)b''(\theta) + c''(\theta)] f(y, \theta) + [a(y)b'(\theta) + c'(\theta)]^2 f(y, \theta) \end{aligned}$$

指数型分布族の性質

$$\begin{aligned} [a(y)b'(\theta) + c'(\theta)]^2 f(y, \theta) &= b'(\theta)^2 \left[a(y) + \frac{c'(\theta)}{b'(\theta)} \right]^2 f(y, \theta) \\ &= b'(\theta)^2 [a(y) - E[a(Y)]]^2 f(y, \theta) \end{aligned}$$

$$\begin{aligned} \int [a(y)b'(\theta) + c'(\theta)]^2 f(y, \theta) dy &= \int b'(\theta)^2 [a(y) - E[a(Y)]]^2 f(y, \theta) dy \\ &= b'(\theta)^2 \int [a(y) - E[a(Y)]]^2 f(y, \theta) dy \\ &= b'(\theta)^2 \text{var}[a(Y)] \end{aligned}$$

$$\begin{aligned} \int [a(y)b''(\theta) + c''(\theta)] f(y, \theta) dy &= \int a(y)b''(\theta) f(y, \theta) dy + \int c''(\theta) f(y, \theta) dy \\ &= b''(\theta) \int a(y) f(y, \theta) dy + c''(\theta) \int f(y, \theta) dy \\ &= b''(\theta) E[a(Y)] + c''(\theta) \end{aligned}$$

$$\frac{d^2}{d\theta^2} \int f(y; \theta) dy = \frac{d^2}{d\theta^2} 1 = 0$$

指数型分布族の性質

$$\begin{aligned} 0 &= b''(\theta)E[a(Y)] + c''(\theta) + b'(\theta)^2 \text{var}[a(Y)] \\ &= b''(\theta)\left\{-\frac{c'(\theta)}{b'(\theta)}\right\} + c''(\theta) + b'(\theta)^2 \text{var}[a(Y)] \end{aligned}$$

$$b'(\theta)^2 \text{var}[a(Y)] = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)$$

$$\text{var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)^3}$$

指数型分布族の対数尤度関数 最尤推定の準備

- 指数型分布族の確率密度関数

$$f(y, \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

- 指数型分布族の尤度関数

$$L(\theta, y) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

- 指数型分布族の対数尤度関数

$$l(\theta, y) = \log L(\theta, y) = a(y)b(\theta) + c(\theta) + d(y)$$