政治学方法論 I
一般化線形モデル

矢内 勇生

神戸大学　法学部/法学研究科

2014 年 12 月 24 日

今日の内容

1. 一般化線形モデル
   - イントロダクション
   - 指数型分布族 (exponential family of distribution)
   - 一般化線形モデル (generalized linear models)
2. ロジットとプロビット
   - ロジット（ロジスティック）回帰とプロビット回帰
3. R による一般化線形モデル
   - glm()
   - 応答変数がカテゴリ変数（三値以上）の場合

線形モデル

- 線形モデル：

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} \\ Y_i &\sim \mathrm{N}(\mu_i, \sigma^2) \end{aligned}$$

- $\mathbf{x}_i^T$ は説明変数行列（計画行列, design matrix）$\mathbf{X}$ の第 i 行
- 一般化線形モデル：線形モデルを拡張する
  1. 応答変数が正規分布以外の分布（離散型分布も含む）に従う場合も扱う
  2. 応答変数と説明変数の関係が線形でない場合も扱う

線形モデルから一般化線形モデルへの拡張

一般化線形モデル (generalized linear models: GLMs)

1. 応答変数が正規分布以外の分布に従う場合も扱う
   - 正規分布は指数型分布族 $\rightarrow$ 指数型分布族に拡張する
2. 応答変数と説明変数の関係が線形でない場合も扱う
   - $\mathrm{E}(Y_i) = \mu_i$ と線形予測子 $\mathbf{x}_i^T \boldsymbol{\beta}$ を線形でない関数 g で結びつける

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \quad \text{or} \quad \mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

   - g を**リンク関数**と呼ぶ

指数型分布族

- 指数型分布族：唯一の母数 θ をもつ確率変数 Y の確率密度（質量）関数が次の形で表されるもの

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

- a, b, s, t は既知の関数
- $s(y) = \exp[d(y)]$, $t(\theta) = \exp[c(\theta)]$
- y と θ が対称
- $a(y) = y$ のもの：正準形 (canonical form)
- $b(\theta)$：自然母数 (natural parameter)
- 注目する母数 θ 以外の母数：撹乱母数 (nuisance parameter)

指数型分布族に属する確率分布

- 正規分布
- ベルヌーイ分布、二項分布
- ポアソン分布
- 負の二項分布
- ベータ分布
- ガンマ分布
- ワイブル分布
- ウィッシャート分布
- ディリクレ分布
- etc.

正規分布 (normal distribution)

- 正規分布の確率密度関数：μ を母数、σ^2 を撹乱母数とする

$$
\begin{aligned}
f(y|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\
&= \exp[\log(2\pi\sigma^2)^{-\frac{1}{2}}] \exp\left[-\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right] \\
&= \exp\left[y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y}{2\sigma^2}\right]
\end{aligned}
$$

- $a(y) = y \to$ 正準形
- 自然母数：$b(\mu) = \mu/\sigma^2$
- $c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$
- $d(y) = -y/2\sigma^2$

二項分布 (binomial distribution)

- 1 回の試行の成功確率が π で、独立な n 回の試行のうち成功する回数を確率変数 Y とする：$Y_i \sim \mathrm{Bin}(n_i, \pi_i)$

$$
\begin{aligned}
f(y|\pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\
&= \exp\left[\log \binom{n}{y} \pi^y (1-\pi)^{n-y}\right] \\
&= \exp\left[y\{\log\pi - \log(1-\pi)\} + n\log(1-\pi) + \log\binom{n}{y}\right]
\end{aligned}
$$

- $a(y) = y \rightarrow$ 正準形
- 自然母数：$b(\pi) = \log\pi - \log(1-\pi) = \log\frac{\pi}{1-\pi}$：ロジット
- $c(\pi) = n\log(1-\pi)$
- $d(y) = \log\binom{n}{y}$

ポアソン分布 (Poisson distribution)

- 決められた時間（または空間）内で特定の事象が起きる回数を確率変数 Y とする：$Y_i \sim \mathrm{Poisson}(\theta_i)$

$$
\begin{aligned}
f(y|\theta) &= \frac{\theta^y \exp(-\theta)}{y!} \\
&= \exp(y \log \theta - \theta - \log y!)
\end{aligned}
$$

- $a(y) = y \to$ 正準形
- 自然母数：$b(\theta) = \log \theta$
- $c(\theta) = -\theta$
- $d(y) = -\log y!$

二項分布とポアソン分布のどちらを使うか

- 二項分布：$X_i \sim \mathrm{Bin}(n_i, \pi_i)$
- ポアソン分布：$Y_i \sim \mathrm{Poisson}(\theta_i)$
- 共通点：どちらも特定の事象が起きる回数を数えている
- 相違点：二項分布の X_i には上限があるが、ポアソン分布の Y_i には上限がない：
  - X_i：0 以上 n_i 以下の整数
  - Y_i：0 以上の整数

  $\rightarrow$ 事象の発生回数とは独立に試行回数 n_i が決められているときは二項分布、そうでなければポアソン分布
- どちらも過分散 (overdispersion) の可能性がある
  - 二項分布で過分散 $\rightarrow$ ベータ二項分布 (beta-binomial)
  - ポアソン分布で過分散 $\rightarrow$ 負の二項分布 (negative binomial)

一般化線形モデルの定義

指数型分布族の確率分布に従う確率変数 $Y_1, \ldots, Y_n$

1. 各 Y_i の確率分布が正準形で母数が 1 つ：

$$f(y_i|\theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)]$$

2. すべての Y_i が同じ確率分布（母数の値は異なってもよい）に従う $\rightarrow$ $Y_1, \ldots, Y_n$ の同時分布：

$$\begin{aligned} & f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n) \\ = & \prod_{i=1}^{n} \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] \\ = & \exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right] \end{aligned}$$

一般化線形モデルの（通常の）目的

- 目的：θ_i の推定ではない $\rightarrow \beta_1, \ldots, \beta_k$ の推定（ただし、$k < n$）
- μ_i を θ_i の関数、$\mathrm{E}(Y_i) = \mu_i$ とし、次の関数 g を考える

$$
\begin{aligned}
g(\mu_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \\
\mu_i &= g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})
\end{aligned}
$$

1. g は単調な関数（単調増加、単調減少、または定数関数）
2. $\mathbf{x}_i^T$ は 1 行 k 列の説明変数ベクトル
3. $\boldsymbol{\beta}$ は k 行 1 列の母数ベクトル

一般化線形モデルの構成要素

1. 同一の確率分布（指数型分布族のもの）に従う応答変数 $Y_1, \ldots, Y_n$
2. 母数ベクトル $\boldsymbol{\beta}$ と説明変数行列 $\mathbf{X}$：

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

3. 単調なリンク関数 g：

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \quad \text{or} \quad \mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

ただし、$\mu_i = \mathrm{E}(Y_i)$

潜在変数を使ったロジスティック回帰の定式化

- 応答変数 Y_i を、連続型の潜在変数 Z_i を使ってモデル化する

$$
\begin{aligned}
y_i &= \begin{cases} 1 & (z_i > 0) \\ 0 & (z_i < 0) \end{cases} \\
z_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i
\end{aligned}
$$

- ただし、ϵ_i はロジスティック分布に従う：

$$\Pr(\epsilon_i < x) = \mathrm{logit}^{-1}(x), \quad \forall x$$

- したがって、

$$\Pr(y_i = 1) = \Pr(z_i > 0) = \Pr(\epsilon_i > -\mathbf{x}_i^T \boldsymbol{\beta}) = \mathrm{logit}^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

プロビット回帰モデル

- 応答変数 Y_i を、連続型の潜在変数 Z_i を使ってモデル化する

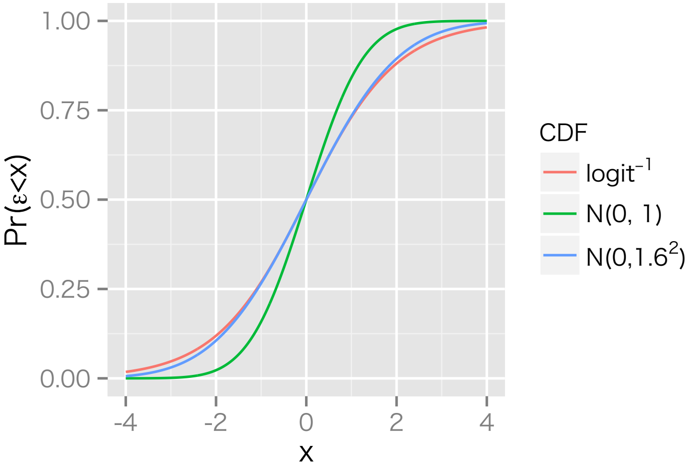
$$
\begin{aligned}
y_i &= \begin{cases} 1 & (z_i > 0) \\ 0 & (z_i < 0) \end{cases} \\
z_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i
\end{aligned}
$$

- ただし、$\epsilon_i \sim \mathrm{N}(0, 1)$
- したがって、

$$
\Pr(y_i = 1) = \Phi(\mathbf{x}_i^T \boldsymbol{\beta})
$$

- ただし、Φ は標準正規分布の累積分布関数 (cdf: cumulative distribution function)

ロジットとプロビットの累積分布関数



ロジットとプロビットの違い

- 潜在変数 Z_i を使った回帰モデル：

$$
\begin{aligned}
y_i &= \begin{cases} 1 & (z_i > 0) \\ 0 & (z_i < 0) \end{cases} \\
z_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i
\end{aligned}
$$

- $\epsilon_i \sim \mathrm{N}(0, 1.6^2)$ とする
- このモデルの推定結果はロジスティック回帰モデルとほぼ一緒
- ロジスティック（ロジット）回帰 $\approx$ プロビット回帰の標準偏差を 1.6 倍したもの

σ を推定できるか？

- 潜在変数を使った定式化で、より一般的に、$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ とし、σ を推定できるか？
- 答え：できない！
- 以下のモデルはすべて等しい：

$$
\begin{aligned}
z_i &= \beta_1 + \beta_2 x_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, 1.6^2) \\
z_i &= (10\beta_1) + (10\beta_2) x_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, 16^2) \\
z_i &= (100\beta_1) + (100\beta_2) x_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, 160^2)
\end{aligned}
$$

- σ を固定する必要 $\rightarrow \sigma = 1$：プロビット（一般化線形モデルの σ は撹乱変数）

glm() で分析できるモデル

- 以下のモデルはすべて glm() で分析できる
    - 線形回帰モデル
    - ロジスティック（ロジット）回帰モデル
    - プロビット回帰モデル
    - ポアソン回帰モデル
    - ベータ二項分布モデル
    - 負の二項分布モデル
- 応答変数がカテゴリー変数の場合：他の関数を使う（後述）

glm() を使うときに特定すべきもの

1. 応答変数ベクトル；$\mathbf{y}$
2. 線形予測子：$\mathbf{X}\boldsymbol{\beta}$
    - 説明変数行列（計画行列）：$\mathbf{X}$
    - 母数ベクトル：$\boldsymbol{\beta}$
3. リンク関数：glm の link を決める
4. 応答変数の確率分布：glm の family を決める
5. 撹乱母数：線形予測子、リンク関数、確率分布に登場する、$\mathbf{X}$ 以外の母数

線形回帰モデル

- リンク関数：恒等関数 (identity function)

$$\mathbf{x}_i^T \boldsymbol{\beta} = g(\mu_i) = \mu_i$$

- 応答変数の確率分布：

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2), \quad \mathrm{E}(Y_i) = \mu_i$$

- family の特定： `family=gaussian(link="identity")`
- 撹乱母数：σ^2

ロジスティック回帰モデル

- リンク関数：ロジット関数 (logit function)

$$\mathbf{x}_i^T \boldsymbol{\beta} = g(\pi_i) = \mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

- 応答変数の確率分布：

$$Y_i \sim \mathrm{Bernoulli}(\pi_i), \quad \mathrm{E}(Y_i) = \pi_i$$

- family の特定： `family=binomial(link="logit")`

プロビット回帰モデル

- リンク関数：プロビット関数 (probit function)

$$\mathbf{x}_i^T \boldsymbol{\beta} = g(\pi_i) = \Phi^{-1}(\pi_i)$$

- 応答変数の確率分布：

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad \text{E}(Y_i) = \pi_i$$

- family の特定： `family=binomial(link="probit")`

ポアソン回帰モデル

- リンク関数：対数関数 (logarithmic function)

$$\mathbf{x}_i^T \boldsymbol{\beta} = g(\theta_i) = \log \theta_i$$

- 応答変数の確率分布：

$$Y_i \sim \mathrm{Poisson}(\theta_i), \quad \mathrm{E}(Y_i) = \theta_i$$

- family の特定： `family=poisson(link="log")`

応答変数がカテゴリ変数のモデル

- 応答変数が順序尺度のとき
  - 順序ロジット回帰 (ordered logit)
  - 順序プロビット回帰 (ordered probit)
- 応答変数が名義尺度のとき
  - 多項（順序なし）ロジット回帰 (multinomial or unordered logit)
  - 多項（順序なし）プロビット回帰 (multinomial or unordered probit)

R での分析法

- 順序ロジット・プロビットは以下の関数で分析可能
  1. MASS パッケージの polr()
  2. arm パッケージの bayespolr()
  3. ordinal パッケージの clm()
- 多項ロジット・プロビットは以下の関数で分析可能
  - ロジット
    1. mlogit パッケージの mlogit()
    2. VGAM パッケージの multinomial()
  - プロビット：MNP パッケージの mnp()