Research Methods in Political Science I

12. Generalized Linear Models

Yuki Yanai

School of Law and Graduate School of Law

January 13, 2016

KOBE UNIVERSITY

Today's Menu

Linear Models

- Linear model

$$E(Y_i) = \mu_i = \mathbf{x}_i^T \beta$$
$$Y_i \sim N(\mu_i, \sigma^2)$$

- $\mathbf{x}_i^T$: the i-th row of the design matrix (predictor matrix) $\boldsymbol{X}$
- Generalized linear models: extensions of linear models
  1. Non-normal response variables (including discrete responses)
  2. Non-linear relationship between the response and the predictors

From Linear Models to Generalized Linear Models

Generalized linear models（GLMs, 一般化線形モデル）

1. Non-normal responses
   - Normal distribution belongs to the exponential family $\rightarrow$ extend to the exponential family
2. Non-linear relationship b/w response and the predictors
   - Link $\mathrm{E}(Y_i) = \mu_i$ to the linear predictor $\boldsymbol{x}_i^T \boldsymbol{\beta}$ by non-linear function g

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

   - g: **Link function**

Exponential Family of Distribution

- Exponential family: PDF (PMF) of a random variable Y with the single parameter θ is represented in the following form

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

- a, b, s, t are some known functions
- $s(y)=\exp[d(y)]$, $t(\theta)=\exp[c(\theta)]$
- y and θ are symmetric
- When $a(y)=y$: Canonical form（正準形）
- $b(\theta)$: Natural parameter（自然母数）
- Parameters other than θ: Nuisance parameter（撹乱母数）

Probability Distributions in the Exponential Family

- Normal
- Bernoulli, Binomial
- Poisson
- Negative binomial
- Beta
- Gamma
- Weibull
- Wishart
- Dirichlet
- etc.

Normal Distribution

- Normal PDF with parameter μ and nuisance parameter σ^2: $Y \sim \mathrm{N}(\mu, \sigma^2)$

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

- $a(y) = y \rightarrow$ Canonical form
- Natural parameter $b(\mu) = \mu/\sigma^2$
- $c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$
- $d(y) = -y/2\sigma^2$

Binomial Distribution

- The number of successes y in n independent Bernoulli trials with success probability π $Y_i \sim \text{Bin}(n_i, \pi_)$

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

- $a(y) = y \rightarrow$ Canonical form
- Natural parameter: $b(\pi) = \log\pi - \log(1-\pi) = \log\frac{\pi}{1-\pi}$: logit
- $c(\pi) = n\log(1-\pi)$
- $d(y) = \log\binom{n}{y}$

Poisson Distribution

- the number of the event occurrences Y in a given time (or space) $Y_i \sim \text{Poisson}(\theta_i)$

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

- $a(y) = y \rightarrow$ Canonical form
- Natural parameter: $b(\theta) = \log\theta$
- $c(\theta) = -\theta$
- $d(y) = -\log y!$

Binomial or Poisson?

- Binomial: $X_i \sim \text{Bin}(n_i, \pi_i)$
- Poisson: $Y_i \sim \text{Poisson}(\theta_i)$
- Common: count the number of events
- Different: Binomial count X_i has the upper limit n_i, but Poisson count Y_i doesn't
    - X_i is an integer in $[0, n_i]$
    - Y_i in a non-negative integer

  $\rightarrow$ Use binomial when n_i is independent of the number of events. Otherwise, use Poisson.
- Both has possibility of overdispersion
    - Binomial with overdispersion $\rightarrow$ Beta-binomial distribution
    - Poisson with overdispersion $\rightarrow$ Negative binomial distribution

Generalized Linear Models

Random variables $Y_1, \ldots, Y_n$ following a distribution of the exponential family

1. PDF (PMF) of each Y_i is in canonical form and has one parameter (save nuisance parameters):

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

2. All Y_i follows the same distribution (θ_i can be different) $\rightarrow$ Joint distribution of $Y_1, \ldots, Y_n$:

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

Purpose of GLMs

- Purpose: Estimating not θ_i but $\beta_1, \ldots, \beta_k$ $(k < n)$
- Suppose μ_i is a function of θ_i and $\mathrm{E}(Y_i) = \mu_i$. Consider the following function g.

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

1. g is a monotonic function (increasing, decreasing, or constant)
2. $\boldsymbol{x}_i^T$ is the $1 \times k$ matrix of the predictors (row vector)
3. $\boldsymbol{\beta}$ is the $k \times 1$ matrix of the parameters (column vector)

Components of GLMs

1. Response following the same distribution (in the exponential family): $Y_1, \ldots, Y_n$
2. Parameter vector β and design matrix $\boldsymbol{X}$:

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

3. Monotonic link function g:

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

where $\mu_i = \mathrm{E}(Y_i)$

Logistic Regression and a Latent Variable

- Model the response Y_i with a continuous latent variable Z_i:

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

- ε_i follows the logistic distribution:

$$\Pr(\varepsilon_i \leq x) = \operatorname{logit}^{-1}(x), \quad \forall x$$

- Therefore,

$$\begin{aligned} \Pr(y_i = 1) &= \Pr(z_i \geq 0) = \Pr(\varepsilon_i \geq -\boldsymbol{x}_i^T \boldsymbol{\beta}) = \Pr(\varepsilon_i \leq \boldsymbol{x}_i^T \boldsymbol{\beta}) \\ &= \operatorname{logit}^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \end{aligned}$$

Probit Model

- Model the responseY_i with a continuous latent variable Z_i:

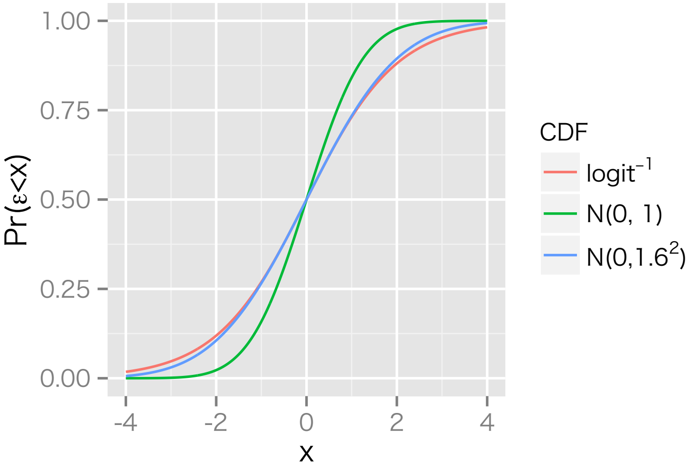
$$
\begin{aligned}
y_i &= \begin{cases} 1 & (z_i \geq 0) \\ 0 & (z_i < 0) \end{cases} \\
z_i &= \mathbf{x}_i^T \beta + \varepsilon_i
\end{aligned}
$$

- $\varepsilon_i \sim \mathrm{N}(0, 1)$
- Therefore,

$$
\begin{aligned}
\Pr(y_i = 1) &= \Pr(z_i \geq 0) = \Pr(\varepsilon_i \geq -\boldsymbol{x}_i^T \boldsymbol{\beta}) = \Pr(\varepsilon_i \leq \boldsymbol{x}_i^T \boldsymbol{\beta}) \\
&= \Phi(\mathbf{x}_i^T \beta)
\end{aligned}
$$

where Φ is the standard normal CDF

CDFs of Logit and Probit



Difference between Logit and Probit

- Regression model with a latent variable Z_i:

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

- Suppose $\varepsilon_i \sim \mathrm{N}(0, 1.6^2)$
- Estimation result of this models is almost same as that of logistic model
- Logistic (logit) $\approx$ Probit with the sd multiplied by 1.6

Can We Estimate σ in a Latent Variable Model?

- Can we set $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ and estimate σ?
- Answer: No!
- Following models are equivalent:

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

- Need fix $\sigma \to \sigma = 1$: Probit
- σ in GLMs is a nuisance parameter

Models Estimated by glm() in R

- Following models are estimated by `glm()`
  - Linear regression
  - Logistic (logit) regression
  - Probit regression
  - Poisson regression
  - Beta-binomial
  - Negative-binomial
- For non-binary categorical response: use different functions in R

What to Specify in glm()

1. Response variable vector: $\boldsymbol{y}$
2. Linear predictor: $\boldsymbol{X\beta}$
   - Design matrix: $\boldsymbol{X}$
   - Parameter vector: $\boldsymbol{\beta}$
3. **Link function**: "link" argument of `glm()`
4. **Probability distribution of the response**: "family" argument of `glm()`
5. Nuisance parameters: parameters other than $\boldsymbol{\beta}$ that appear in the link function or the distribution

Linear Regression Model

- **Link**: Identity function（恒等関数）

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

- **Probability distribution of the response**:

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

- Specifying "family" of `glm()`:
  `family = gaussian(link = "identity")`
- Nuisance parameter: σ^2

Logistic Regression Model

- **Link**: Logit function

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

- **Probability distribution of the response**:

$$Y_i \sim \text{Bernoulli}(\pi_i) = \text{Binomial}(n=1, \pi_i), \quad \mathrm{E}(Y_i) = \pi_i$$

- Specifying “family” of `glm()`:
`family = binomial(link = "logit")`

Probit Regression Model

- **Link**: Probit function

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

- **Probability distribution of the response**:

$$Y_i \sim \text{Bernoulli}(\pi_i) = \text{Binomial}(n = 1, \pi_i), \quad \mathrm{E}(Y_i) = \pi_i$$

- Specifying "family" of `glm()`:
  `family = binomial(link = qy"probit")`

Poisson Regression Model

- **Link**: Logarithmic function

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

- **Probability distribution of the response**:

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

- Specifying "family" of `glm()`:
`family = poisson(link = "log")`

Models for Non-binary Categorical Responses

- Ordinal response
  - Ordered logit
  - Ordered probit
- Nominal response
  - Multinomial (unordered) logit
  - Multinomial (unordered) probit)

Some R Functions

- Ordered logit or probit
  1. MASS::polr()
  2. arm::bayespolr()
  3. ordinal::clm()
- Multinomial models
  - Multinomial logit
    1. mlogit::mlogit()
    2. VGAM::multinomial()
  - Multinomial probit: MNP::mnp()