

GENERALIZED LINEAR MODEL: Lecture I: Introduction

Hannu Oja

Spring 2006

- Dobson, A. J. (1990), *An Introduction to Generalized Linear Models*, Chapman and Hall
- McCullagh, P. and Nelder, J.A. (1986), *Generalized Linear Models*, Chapman and Hall.

Plan for Lecture I

Overview on model construction for different types of data:

1. Data matrix, response variable, explaining variables
2. Continuous response variable - normal distribution
3. Dichotomous response variable - Bernoulli distribution
4. Count response variable - Binomial and Poisson distribution
5. Life time response - Gamma distribution
6. Exponential family of distribution
7. Generalized linear model
8. Likelihood inference

Data: Response, explaining variables

- Observed data matrix (rows=individuals, columns= variables, x -variables are explaining variables, y is the response variable)

$$(\mathbf{X}, \mathbf{y}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} & y_1 \\ x_{21} & x_{22} & \dots & x_{2p} & y_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} & y_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_1 & y_1 \\ \mathbf{x}'_2 & y_2 \\ \dots & \dots \\ \mathbf{x}'_n & y_n \end{pmatrix}$$

- **General assumptions:** Design matrix \mathbf{X} is assumed to be fixed. The response measurements y_1, \dots, y_n are independent; the probability distribution of y_i depends on the vector \mathbf{x}_i .
- **General aim:** Explore the effect of \mathbf{x} on the distribution of response variable y
- Types of the response: Continuous, positive continuous, dichotomous, frequency data (counts)
- Types of the explaining variables: Continuous, dichotomous, categorical

Continuous response - normal distribution

- Continuous random variable y has a normal distribution with mean value μ and variance σ^2 , write $y \sim N(\mu, \sigma^2)$, if it has a density function

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}}$$

- **Regular linear model:** Assume that y_1, \dots, y_n are independent and

$$y_i \sim N(\mu_i, \sigma^2), \quad \text{where } \mu_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

- Then $\eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ is called **linear predictor** for the i th individual, and β_1, \dots, β_p are **regression coefficients**
- We wish to make inference (tests, estimates, confidence intervals) on unknown parameters β_1, \dots, β_p ; the unknown parameter σ^2 is a **nuisance parameter**.

Dichotomous response - Bernoulli distribution

- Random variable y has a Bernoulli distribution with parameter p , $0 < p < 1$, write $y \sim \text{Ber}(p)$, if y has two possible values 0 and 1 with probabilities

$$P(y = 0) = 1 - p \quad \text{and} \quad P(y = 1) = p.$$

- **Logistic model:** The response variables y_1, \dots, y_n are assumed to be independent, and

$$y_i \sim \text{Ber}(p_i) \quad \text{where} \quad p_i = \frac{e^{\beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

- Again, $\eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ is called **linear predictor**, and expected value $\mu_i = E(y_i)$ is

$$\mu_i = g(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

- We wish to make inference (tests, estimates, confidence intervals) on unknown parameters β_1, \dots, β_p .

Count response - Binomial distribution

- Random variable y has a Binomial distribution with parameters n (positive integer) and p , $0 < p < 1$, (expected value is $\mu = np$), write $y \sim \text{Bin}(n, p)$, if y has values $0, 1, 2, \dots, n$ with probabilities

$$P(y = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots$$

- If y_1, \dots, y_n is a random sample from $\text{Ber}(p)$, then $y = \sum_{i=1}^n y_i \sim \text{Bin}(n, p)$.
- Often, the observed data matrix with Bernoulli responses is grouped as

$$(\mathbf{X}, \mathbf{n}, \mathbf{y}) = \begin{pmatrix} \mathbf{x}'_1 & n_1 & y_1 \\ \mathbf{x}'_2 & n_2 & y_2 \\ \dots & \dots & \dots \\ \mathbf{x}'_r & n_r & y_r \end{pmatrix}$$

- **Logistic model:** The response variables y_1, \dots, y_n are assumed to be independent, and

$$y_i \sim \text{Bin}(n_i, p_i) \quad \text{where} \quad p_i = \frac{e^{\beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

Count response - Poisson distribution

- Random variable y has a Poisson distribution with expected value μ , write $y \sim Poi(\mu)$, if y has values $0, 1, 2, \dots$ with probabilities

$$P(y = k) = \frac{\mu^k}{k!} e^{-\mu}, \quad k = 0, 1, 2, \dots$$

- **Log-linear model:** The response variables y_1, \dots, y_n are assumed to be independent, and

$$y_i \sim Poi(\mu_i) \quad \text{where} \quad \mu_i = e^{\beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

- Again, $\eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ is called **linear predictor**, and expected value $\mu_i = E(y_i)$ is

$$\mu_i = g(\eta_i) = e^{\eta_i}$$

(Thus $\log(\mu_i) = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ - log of expected value is linear.)

- We wish to make inference (tests, estimates, confidence intervals) on unknown parameters β_1, \dots, β_p .

Continuous positive response - Gamma distribution

- Continuous positive random variable y has a gamma distribution with mean value $\mu > 0$ and shape (nuisance) parameter $\nu > 0$, write $y \sim \text{Gamma}(\mu, \nu)$, if it has a density function

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} \exp\left\{-\frac{\nu y}{\mu}\right\}$$

- **Gamma model:** Assume that y_1, \dots, y_n are independent and

$$y_i \sim \text{Gamma}(\mu_i, \nu), \quad \text{where } \mu_i = e^{\beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

- Then $\eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ is called **linear predictor** for the i th individual, and β_1, \dots, β_p are **regression coefficients**
- We wish to make inference (tests, estimates, confidence intervals) on unknown parameters β_1, \dots, β_p ; the shape parameter ν is a **nuisance parameter**.

Exponential family of distributions

- Random variable (continuous, dichotomous, count) has a distribution belonging to the exponential family, write $y \sim E(\theta, \phi)$ if the probability or density function is of the form

$$f(y) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

- All the distributions above (normal, Bernoulli, binomial, Poisson, gamma) belong to this family (left as an exercise)
- Parameter θ is called **canonical parameter**; ϕ is nuisance. The function $b(\theta)$ can be used to find the expected value and variance as functions of θ :

$$\mu = E(y) = b'(\theta) \quad \text{and} \quad \text{Var}(y) = b''(\theta)a(\phi).$$

Generalized linear model

- Assume that y_1, \dots, y_n are independent and the distribution of y_i belongs to an exponential family with known functions a , b , and c but unknown θ_i and ϕ . Then $\mu_i = E(y_i) = b'(\theta_i)$.
- $\eta_i = \beta^T \mathbf{x} = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ is called **linear predictor** for the i th individual, and $\beta = (\beta_1, \dots, \beta_p)^T$ is the vector of **regression coefficients**
- Link connecting the distribution and linear predictor is

$$g(\mu_i) = \eta_i.$$

Function $g(\mu)$ is called the **link function**.

- We wish to make inference (tests, estimates, confidence intervals) on unknown parameters β_1, \dots, β_p ; the unknown parameter ϕ is a **nuisance parameter**.

Likelihood inference

- Let the density or probability function of y_i be $f_{\mu_i, \phi}(y)$ where $g(\mu_i) = \eta_i = \beta^T \mathbf{x}_i$. The **likelihood function** corresponding to data set (\mathbf{X}, \mathbf{y}) is defined as

$$L(\beta, \phi) = \prod_{i=1}^n f_{\mu_i, \phi}(y_i)$$

- Rough interpretation: $L(\beta, \psi)$ gives the "probability" that parameter values (β, ψ) produces the observed values in \mathbf{y} .
- The parameter value $(\hat{\beta}, \hat{\psi})$ which maximizes likelihood function is called the **maximum likelihood estimate** of (β, ψ) .
- It is often easier to work with the **log-likelihood function**

$$l(\beta, \phi) = \log L(\beta, \phi) = \sum_{i=1}^n \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi) \right\}$$