

STAT347: Generalized Linear Models

Lecture 2

Today's topics: Agresti Chapters 4.1-4.2

- The exponential dispersion family
- Likelihood score equations for parameter estimation

1 The exponential dispersion family

1.1 Definition

The observation y_i follows an exponential dispersion family distribution and has the density $f(y_i; \theta_i, \phi)$ of the form ("density" here including the possibility of discrete atoms.)

$$f(y_i; \theta_i, \phi) = e^{\frac{y_i \theta_i - b(\theta_i)}{a(\phi)}} f_0(y_i; \phi)$$

Terminologies:

- θ : natural or canonical parameters
- $b(\theta)$: normalizing or cumulant function
- ϕ : dispersion parameter with $a(\phi) > 0$
- Typically $a(\phi) \equiv 1$ and $f_0(y; \phi) = f_0(y)$. An exception is the Gaussian distribution where $a(\phi) = \sigma^2$

1.2 Some well-known one-parameter exponential families

1. Normal with mean μ_i and variance σ^2 :

$$f(y_i; \mu_i, \sigma) = e^{\frac{y_i \mu_i - \mu_i^2/2}{\sigma^2}} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y_i^2}{2\sigma^2}} \right]$$

2. Bernoulli with probability p_i :

$$\begin{aligned} f(y_i; p_i) &= p_i^{y_i} (1 - p_i)^{1-y_i} = e^{y_i \log \frac{p_i}{1-p_i} + \log(1-p_i)} \\ &= e^{y_i \theta_i - \log[1+e^{\theta_i}]} \end{aligned}$$

3. Binomial with p_i and n_i :

$$\begin{aligned} f(y_i; p_i, n_i) &= \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} = e^{y_i \log \frac{p_i}{1-p_i} + n_i \log(1-p_i)} \binom{n_i}{y_i} \\ &= e^{y_i \theta_i - n_i \log[1+e^{\theta_i}]} \binom{n_i}{y_i} \end{aligned}$$

4. Poisson with mean λ_i :

$$f(y_i; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = e^{y_i \log \lambda_i - \lambda_i} \frac{1}{y_i!} = e^{y_i \theta_i - e^{\theta_i}} \frac{1}{y_i!}$$

1.3 Moment relationships

Take the first and second derivative respect to θ_i for both sides of the equation

$$e^{b(\theta_i)/a(\phi)} = \int e^{y_i \theta_i / a(\phi)} f_0(y_i; \phi) dy_i$$

We can derive:

$$\begin{aligned} \mu_i &= \mathbb{E}(y_i) = b'(\theta_i) \\ V_{\theta_i} &= \text{Var}(y_i) = b''(\theta_i) a(\phi) \end{aligned}$$

In addition, this indicates that:

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\text{Var}(y_i)}{a(\phi)} > 0$$

thus the mapping from θ_i to μ_i is one to one increasing.

1.4 The canonical link function in GLM

Assume

$$g(\mu_i) = \theta_i = X_i^T \beta$$

(Why? Easier calculations)

As $\mu_i = b'(\theta_i)$, the link function will be

$$g(\cdot) = (b')^{-1}(\cdot)$$

which is called the canonical link.

Canonical link functions for Binomial, Poisson and Bernoulli distributions.

2 Likelihood score equations

Assume each observation y_i follows an exponential dispersion distribution

$$f(y_i; \theta_i, \phi) = e^{\frac{y_i \theta_i - b(\theta_i)}{a(\phi)}} f_0(y_i; \phi)$$

and the link function $g(\mu_i) = X_i^T \beta$. Then for n independent observations, the log likelihood is

$$L = \sum_i L_i = \sum_i \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_i \log f_0(y_i; \phi)$$

2.1 For the canonical link

If $g(\mu_i) = \theta_i = X_i^T \beta$, then

$$L = \frac{1}{a(\phi)} \left[\sum_j \left(\sum_i y_i x_{ij} \right) \beta_j - \sum_i b(X_i^T \beta) \right] + \sum_i \log f_0(y_i; \phi)$$

- Score equation for β_j

$$\frac{\partial L}{\partial \beta_j} = \frac{1}{a(\phi)} \left[\sum_i y_i x_{ij} - \sum_i b'(X_i^T \beta) x_{ij} \right] = \frac{1}{a(\phi)} \left[\sum_i (y_i - \mu_i) x_{ij} \right] = 0$$

which is equivalent to

$$\sum_i (y_i - \mu_i) x_{ij} = 0$$

- score equation for a Poisson and Gaussian canonical link model (Section 4.2.2)

Gaussian model:

$$\sum_i (y_i - X_i^T \beta) x_{ij} = 0$$

Poisson model:

$$\sum_i (y_i - e^{X_i^T \beta}) x_{ij} = 0$$

- L is a concave function of β :

$$\frac{\partial}{\partial \beta} \left[\sum_i (y_i - \mu_i) X_i \right] = - \sum_i \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta} X_i^T = - \sum_i \frac{\text{Var}(y_i)}{a(\phi)} X_i X_i^T \prec 0$$

2.2 For a general link

Let $\eta_i = g(\mu_i) = X_i^T \beta$ Then

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

We have

- $\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$
- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{\text{Var}(y_i)}$
- $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial g(\mu_i)} = \frac{1}{g'(\mu_i)}$
- $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

Thus, the score equation

$$\frac{\partial L}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(y_i)} \frac{1}{g'(\mu_i)} = 0$$

- The score equation only depends on the mean and variance of y
- Matrix form of the score equation:

$$\dot{L}(\beta) = X^T D V^{-1} (y - \mu) = 0$$

where $V = \text{diag}(\text{Var}(y_1), \dots, \text{Var}(y_n))$ and $D = \text{diag}(g'(\mu_1), \dots, g'(\mu_n))^{-1}$, $y = (y_1, \dots, y_n)$ and $\mu = (\mu_1, \dots, \mu_n)$.

- L is not necessarily a concave function of β .

- A special case: if g is the canonical link, then $D = \frac{1}{a(\phi)} V$ (as $g'(\mu_i) = \partial\theta_i/\partial\mu_i$), thus the score equation is simplified to

$$\frac{1}{a(\phi)} X^T (y - \mu) = 0$$

the same as what we derived earlier.

- Another special case: if we assume $g(\mu_i) = \mu_i = X_i^T \beta$, then the estimation equations become

$$\sum_i \frac{(y_i - X_i^T \beta) X_i}{\text{Var}(y_i)} = 0$$

3 Asymptotic distribution of GLM

- the MLE $\hat{\beta}$ is consistent when $n \rightarrow \infty$ and p is fixed.
- Asymptotic normality: when n is large

$$\hat{\beta} - \beta_0 \sim N(0, V_{\beta_0})$$

where β_0 is the true value of the parameter. ($nV_{\beta_0} = O(1)$)

As an applied course, we ignore the discussions of the conditions of the above consistency and CLT results, and also skip the proofs.

3.1 Calculation of V_{β_0}

Delta method:

$$0 = \dot{L}(\hat{\beta}) \approx \dot{L}(\beta_0) + \ddot{L}(\beta_0)(\hat{\beta} - \beta_0)$$

The above approximation is a general approach and can be applied to any estimation equation that results in a consistent estimate of β .

Thus

$$\hat{\beta} - \beta_0 \approx - \left(\ddot{L}(\beta_0) \right)^{-1} \dot{L}(\beta_0)$$

- Under appropriate conditions, we have

$$\ddot{L}(\beta_0)/n = \sum_i \ddot{L}_i(\beta_0)/n \rightarrow \text{Const.} \quad (\text{law of large numbers})$$

$$\frac{\dot{L}(\beta_0)}{\sqrt{n}} = \frac{\sum_i \dot{L}_i(\beta_0)}{\sqrt{n}} \xrightarrow{d} N(0, V) \quad (\text{central limit theorem})$$

Thus we have

$$V_{\beta_0} \approx \left(\ddot{L}(\beta_0) \right)^{-1} \text{Var} \left(\dot{L}(\beta_0) \right) \left(\ddot{L}(\beta_0) \right)^{-1}$$

- property of the likelihood:

$$\text{Var} \left(\dot{L}(\beta_0) \right) = \mathbb{E} \left(\left(\frac{\partial L}{\partial \beta} \Big|_{\beta=\beta_0} \right)^2 \right) = -\mathbb{E} \left(\ddot{L}(\beta_0) \right)$$

- $V_{\beta_0} = -\mathbb{E} \left(\ddot{L}(\beta_0) \right)^{-1}$
- $\hat{\beta}$ is more precise when $L(\beta)$ has larger curvature at β_0 .
- See Chapter 4.2.4. $V_{\beta_0} = (X^T W X)^{-1}$ where $W = D^2 V^{-1}$

3.2 The distribution of any function $h(\hat{\beta})$

- $h(\hat{\beta})$ is a consistent estimator of $h(\beta_0)$
- Delta method:

$$h(\hat{\beta}) \approx h(\beta_0) + \dot{h}(\beta_0)^T (\hat{\beta} - \beta_0)$$

$$\sqrt{n} \left(h(\hat{\beta}) - h(\beta_0) \right) \rightarrow N \left(0, n \dot{h}(\beta_0)^T V_{\beta_0} \dot{h}(\beta_0) \right)$$

- Example: fitted values $h_i(\hat{\beta}) = g^{-1}(X_i^T \hat{\beta})$

Next time: Chapter 4.3-4.4, Hypothesis testing, deviance