

# STAT347: Generalized Linear Models

## Lecture 5

Today's topics: Chapter 4.5, 5.1

- GLM computation
- Binary data model: data input, link function
- R data example

### 1 Computation

Let us discuss the case of  $a(\phi) = 1$  to simplify notation. As  $\phi$  does not affect the point estimate of  $\beta$ , when  $a(\phi)$  is not a constant, one can get  $\hat{\beta}$  from the score equations first. Then one can estimate  $\phi$  from MLE with  $\hat{\beta}$  plugged in.

Score equation:

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

where

$$L(\beta) = \sum_i [y_i \theta_i - b(\theta_i)]$$

(This is the log-likelihood ignoring the term involving  $\phi$  that does not affect the estimation of  $\beta$ )

#### 1.1 Newton's method

Second-order approximation of  $L(\beta)$

$$L(\beta) \approx L(\beta^{(t)}) + \dot{L}(\beta^{(t)})^T (\beta - \beta^{(t)}) + \frac{1}{2} (\beta - \beta^{(t)})^T \ddot{L}(\beta^{(t)}) (\beta - \beta^{(t)})$$

at  $t$ th iteration. If  $\ddot{L}(\beta^{(t)}) \preceq 0$ , then maximizing the second-order approximation is equivalent to solving

$$\dot{L}(\beta) \approx \dot{L}(\beta^{(t)}) + \ddot{L}(\beta^{(t)}) (\beta - \beta^{(t)}) = 0$$

We have

$$\beta^{(t+1)} = \beta^{(t)} - \ddot{L}(\beta^{(t)})^{-1} \dot{L}(\beta^{(t)})$$

- Newton's method is a general algorithm for optimizing twice-differentiable functions.
- Generally converge to the global maximum if  $L(\beta)$  is strongly concave
  - If  $g(\cdot)$  is the canonical link, then  $L(\beta)$  is concave in  $\beta$

$$-\ddot{L}(\beta^{(t)}) = X^T W^{(t)} X = \frac{1}{a(\phi)} X^T V^{(t)} X = -\mathbb{E} \left( \ddot{L}(\beta^{(t)}) \right) \succeq 0$$

We showed the first equality in section 2.1 of lecture 2. This shows that the though observed log-likelihood function  $L(\beta)$  is random, its hessian is a constant.

- If  $g(\cdot)$  is a general link, then  $L(\beta)$  is NOT guaranteed to be concave in  $\beta$
- If  $-\ddot{L}(\beta^{(t)})$  is not non-negative, then step  $t$  does not maximize the quadratic approximation and Newton's method may not converge.
- We can use another quadratic approximation that works better in practice: Fisher scoring method

## 1.2 Fisher scoring method

In lecture 2, we showed that  $-\mathbb{E}(\ddot{L}(\beta)) \succeq 0$  for any  $\beta$ .

Instead of using the Hessian  $\ddot{L}(\beta^{(t)})$ , use its expectation

$$J^{(t)} = \mathbb{E}(\ddot{L}(\beta^{(t)})) = -X^T W^{(t)} X$$

instead of  $\ddot{L}(\beta^{(t)})$  itself in the second-order approximation. Each iteration becomes:

$$\beta^{(t+1)} = \beta^{(t)} - \left(J^{(t)}\right)^{-1} \dot{L}(\beta^{(t)})$$

## 1.3 Iteratively reweighted least squares (IRLS)

We can make a connection between the optimization for GLM and weighted least squares estimation.

Recall 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  $\mu = (\mu_1, \dots, \mu_n)$ .

Also in lecture 2, we used the notation  $\eta_i = X_i^T \beta = g(\mu_i)$ . Thus,  $D = \text{diag}\left(\frac{\partial \mu_1}{\partial \eta_1}, \dots, \frac{\partial \mu_n}{\partial \eta_n}\right)$ . We also defined the diagonal matrix  $W = D^2 V^{-1}$ . Thus,

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

We can make a first order approximation of  $\mu$

$$\mu = \mu^{(t)} + D^{(t)} (\eta - \eta^{(t)})$$

then

$$\dot{L}(\beta) \approx X^T W^{(t)} (z^{(t)} - X\beta)$$

where

$$z^{(t)} = X\beta^{(t)} + \left(D^{(t)}\right)^{-1} (y - \mu^{(t)})$$

is a linear approximation of  $\eta$  at the  $t$ th iteration.

Thus, at the  $t + 1$ th iteration, we solve

$$X^T W^{(t)} (z^{(t)} - X\beta) = 0$$

which can be considered as a weighted linear regression with observations  $z_i^{(t)}$  and weight  $w_i$  for each sample  $i$ .

- IRLS is equivalent to Fisher scoring. The  $t$ th step of Fisher scoring satisfy

$$\begin{aligned}(X^T W^{(t)} X) \beta^{(t+1)} &= X^T W^{(t)} X \beta^{(t)} + X^T D^{(t)} (V^{(t)})^{-1} (y - \mu^{(t)}) \\ &= X^T W^{(t)} \left[ X \beta^{(t)} + (D^{(t)})^{-1} (y - \mu^{(t)}) \right] \\ &= X^T W^{(t)} z^{(t)}\end{aligned}$$

- weight matrix  $W^{(t)} \approx \text{Var}(z^{(t)})^{-1}$

## 2 Binary/binomial data model

If the observation  $y_i$  is binomial

$$y_i \sim \text{Binomial}(n_i, p_i)$$

and probability function:

$$f(y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} = \binom{n_i}{y_i} \left( \frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i)^{n_i}$$

If  $n_i = 1$ , then  $y_i$  is a 0/1 binary data point (follows a Bernoulli distribution).

### 2.1 Data input

If  $X_i$  are categorical variables, then we may have samples with the same  $X_i$  and we can group them together.

- ungrouped data: each  $n_i = 1$  and some samples have the same  $X_i$ , thus they share the same  $p_i$
- a grouped sample  $\tilde{y}_k$  for group  $k$  contains  $n_k$  ungrouped samples whose  $X_i$  are the same we only have group level covariates). As  $p_i = g^{-1}(X_i^T \beta)$ , samples within the same group share the same mean. Let  $I_k = \{i : i \text{ in group } k\}$  be the set of individual binary samples and let  $n_k = |I_k|$ . Then the response for the group samples is:

$$\tilde{y}_k = \sum_{i \in I_k} y_i \sim \text{Binomial}(n_k, p_k)$$

- The grouped data follows the Binomial distribution because we assume that the samples are independent within each group.
- If there are some unmeasured group-level covariates that affect all samples in the group, it can bring in extra dependency and an inflated variance of  $\tilde{y}_k$ . (we will discuss this issue later in detail in Chapter 8 and 9.)
- Let  $N = \sum_k n_k$  The likelihood for the ungrouped data is:

$$\begin{aligned}f(y_1, y_2, \dots, y_N) &= \prod_i p_i^{y_i} (1 - p_i)^{1 - y_i} \\ &= \prod_k \prod_{i \in I_k} p_k^{\tilde{y}_k} (1 - p_k)^{n_k - \tilde{y}_k}\end{aligned}$$

The likelihood for the corresponding grouped data is:

$$f(\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_K) = \prod_k \binom{n_k}{y_k} \prod_{i \in I_k} p_k^{\tilde{y}_k} (1 - p_k)^{n_k - \tilde{y}_k}$$

The likelihood is not the same between the grouped data and ungrouped data. However, the log-likelihood function only differs by a constant, thus the GLM solution does not change.

## 2.2 Link function

The expectation of each sample is  $\mathbb{E}(y_i) = n_i p_i$  where  $n_i$  is a known constant. Thus we define the link function as a function of  $p_i$

$$g(p_i) = X_i^T \beta$$

Equivalently,

$$p_i = g^{-1}(X_i^T \beta) \in [0, 1]$$

If  $g$  is a one-to-one mapping (otherwise there can be identifiability issues) and continuous function, then  $g^{-1}$  should be monotone. In that case, one natural choice of  $g^{-1}$  is to make it as a cdf of some distribution. We then can denote  $F(z) = g^{-1}(z)$  as some cdf function. Let  $\epsilon_i \stackrel{i.i.d.}{\sim} F(\cdot)$

$$p_i = F(X_i^T \beta) = \mathbb{P}(\epsilon_i \leq X_i^T \beta) = \mathbb{P}(X_i^T \beta - \epsilon_i \geq 0)$$

If  $y_i$  is binary, this indicates that  $y_i$  follows the distribution

$$Y_i = \begin{cases} 1 & \text{if } X_i^T \beta - \epsilon_i \geq 0 \\ 0 & \text{else} \end{cases}$$

This is also called a latent variable threshold model.

Popular latent variable threshold models:

- The probit link:  $F(z)$  is the cdf of a standard Gaussian distribution

$$p_i = \mathbb{P}(X_i^T \beta - \epsilon_i \geq 0) = \mathbb{P}(X_i^T \beta + \epsilon_i \geq 0)$$

where  $\epsilon_i \sim N(0, 1)$ . Let the hidden variable be  $y_i^* = X_i^T \beta + \epsilon_i$ , then it goes to the definition of the probit link that some of you may be more familiar with:

$$Y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{else} \end{cases}$$

- The logit link:  $F(z)$  is the cdf of a standard logistic distribution

$$F(z) = \frac{e^z}{1 + e^z}$$

- The link function is called the logit link:  $g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$
- The logit link is the canonical link of the Binomial distribution

- The identity link:  $F(z)$  is the cdf of a uniform  $[0, 1]$  distribution and  $p_i = X_i^T \beta$

- The identity link corresponds to a uniform cdf only when  $X_i^T \beta \in [0, 1]$  for all samples.
- Because of the range issue, when using R to solve a binomial GLM with identity link, there can often be numerical problems (such as the error we saw in the earlier data example in Section 1.4, Data Example 1).
- The log-log link:  $F(z)$  is the cdf of a standard double-exponential distribution (Gumbel distribution)

$$F(z) = e^{-e^{-z}}$$

- The link function is called the log-log link:

$$g(p_i) = -\log[-\log(p_i)] = X_i^T \beta$$

- Both the probit and logit link assumes a symmetric  $\epsilon_i$  (around 0). So we implicitly assumed that the response curve is symmetric at 0.5

$$g(p_i) = -g(1 - p_i)$$

One can use the log-log link if such assumption is severely violated (or use a complementary log-log link depending on the shape of the response curve). Read Chapter 5.6.3 for more details (also discussed how one may choose an appropriate link function in practice).

### 3 Data example

Chapter 5.7. Please check the R notebook 3-1.

Next time: Chapter 5.2, 5.3 - 5.5, 5.7, binary GLM: some applications, inference and model fitting