STAT347: Generalized Linear Models
Lecture 4

Today's topics: Chapters 4.5, 4.7

- Computation of the ML estimate
- Example: building a GLM

1 Computation

Log-likelihood:

$$L(\beta) = \sum_i [y_i\theta_i - b(\theta_i)] + \sum_i \log f_0(y_i)$$

Score equation:

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

1.1 Newton's method

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

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

at tth 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.
- Converge to the global maximum if $L(\beta)$ is strongly concave
  - If $g(\cdot)$ is the canonical link, then $L(\beta)$ is concave in β

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

  - If $g(\cdot)$ is a general link, then $L(\beta)$ is NOT guaranteed to be concave in β
  - If $-\ddot{L}(\beta^{(t)})$ is not non-negative, than step i 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 $-\mathbb{E}\left(\ddot{L}(\beta)\right) \succeq 0$ for any β.

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

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

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)

Recall the score equation:

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

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

Also in lecture 2, we used the notation $\eta_i = X_i^T \beta = g(\mu_i)$. Thus, $D = \text{diag}\left(\frac{\partial \mu_1}{\partial \eta_1}, \cdots, \frac{\partial \mu_n}{\partial \eta_n}\right)$. We also defined the diagnoal 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 μ

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

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 η at the tth 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, see Section 4.5.4
- weight matrix $W^{(t)} \approx \text{Var}\left(z^{(t)}\right)^{-1}$

Next time: Chapter 5.1 - 5.2, binary data model, application scenarios