ST509 Computational Statistics

Lecture 4: Generalized Linear Models

Seung Jun Shin

Department of Statistics
Korea University

E-mail: sjshin@korea.ac.kr

Introduction I

- Statistical problem can be viewed as an optimization. (eq, MLE, M-estimation, etc.)
- Optimization often can be rewritten as solving equations. (eq, normal equations)
- Some problems do not have a explicit solution and a numerical approach should be exploited.

Introduction II

- One basic root finding algorithm is the bisection method.
- Suppose $f(x)$ is continuous on $x = [a, b]$ with $f(a)f(b) < 0$.
- Bisection method:
  1. Initialize $l^{(1)} = a$ and $u^{(1)} = b$:
  2. Compute a middle point $m^{(t)} = (l^{(t)} + u^{(1)})/2$ and $f(m^{(t)})$.
  3. Update:
     - $l^{(t+1)} = l^{(t)}$ and $u^{(t+1)} = m^{(t)}$, if $f(l^{(t)})f(m^{(t)}) < 0$.
     - $l^{(t+1)} = m^{(t)}$ and $u^{(t+1)} = u^{(t)}$, if $f(m^{(t)})f(u^{(t)}) < 0$.
  4. Repeat 2-3 until $|u^{(t+1)} - l^{(t+1)}| < \delta$.
  5. The solution is

$$x^* = \frac{u^{(t+1)} - l^{(t+1)}}{2}.$$

Newton-Raphson Method I

- Consider a root finding problem for a continuous and differentiable function f.
$$f(x) = 0$$
- Instead of solving $f(x) = 0$ directly, tackle its linear approximation (i.e., 1st order Tayler expansion).
- For a given x_0, Newton Raphson method solves
$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) = 0$$
and yields
$$x = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Newton-Raphson Method II

1. Initialize $x^{(1)} = x_0$ which can be arbitrary on the domain of $f(x)$.
2. Update for $t = 1, 2, \cdots$

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}$$

until

$$\frac{|x^{(t+1)} - x^{(t)}|}{|x^{(t)}|} < \delta$$

for a small $\delta > 0$.

Algorithm 1: Newton-Raphson method for finding a root

Newton-Raphson Method III

- The idea can naturally be extended to the optimization:

$$x^* := \underset{x}{\operatorname{argmin}}\, f(x)$$

- Direct optimization of $f(x)$ is often difficult. Let's tackle its quadratic approximation (i.e., 2nd order Tayler expansion).
- For a given x_0, we can minimize

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

  with respect to x.
- Taking derivative, we have

$$f'(x_0) + f''(x_0)(x - x_0) = 0$$

  which yields

$$x = x_0 - \frac{f'(x_0)}{f''(x_0)}.$$

Newton-Raphson Method IV

- For multivariate $\mathbf{x} \in \mathbb{R}^p$, we have

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

where

$$\begin{aligned} \text{(Gradient)} \quad \nabla f(\mathbf{x}_0) &= \partial f(\mathbf{x})/\partial \mathbf{x}\Big|_{\mathbf{x}=\mathbf{x}_0}; \\ \text{(Hessian)} \quad \mathbf{H}(\mathbf{x}_0) &= \partial f(\mathbf{x})/\partial \mathbf{x}\partial \mathbf{x}^T\Big|_{\mathbf{x}=\mathbf{x}_0}. \end{aligned}$$

- The updating equation is

$$\mathbf{x} = \mathbf{x}_0 - \mathbf{H}^{-1}(\mathbf{x}_0)\nabla f(\mathbf{x}_0)$$

Newton-Raphson Method V

1. Initialize $\mathbf{x}^{(1)} = \mathbf{x}_0$ which can be arbitrary on the domain of $f(\mathbf{x})$.
2. Update for $t = 1, 2, \cdots$

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \mathbf{H}^{-1}(\mathbf{x}^{(t)})\nabla f(\mathbf{x}^{(t)})$$

   until

$$\frac{\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|}{\|\mathbf{x}^{(t)}\|} < \delta$$

   for a small $\delta > 0$.

Algorithm 2: Newton-Raphson Method for Optimization

Logistic Regression I

- Logistic regression assumes

$$Y \mid \mathbf{x} \sim \text{Bern}\{p(\mathbf{x};\boldsymbol{\beta})\}$$

where

$$\text{logit}\{p(\mathbf{x};\boldsymbol{\beta})\} := \log\left\{\frac{p(\mathbf{x};\boldsymbol{\beta})}{1-p(\mathbf{x}_i;\boldsymbol{\beta})}\right\} = \boldsymbol{\beta}^T\mathbf{x}.$$

or equivalently

$$p(\mathbf{x};\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\boldsymbol{\beta}^T\mathbf{x})}.$$

Logistic Regression II

- Given a set of data $(y_i, \mathbf{x}_i), i = 1, \cdots, n$, the likelihood is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\beta})^{y_i} \{1 - p(\mathbf{x}_i; \boldsymbol{\beta})\}^{1-y_i}$$

- Taking log, we have

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [y_i \log p(\mathbf{x}_i; \boldsymbol{\beta}) + (1 - y_i) \log\{1 - p(\mathbf{x}_i; \boldsymbol{\beta})\}] \\
&= \sum_{i=1}^{n} \left[y_i(\boldsymbol{\beta}^T \mathbf{x}_i) - \log\{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)\} \right]
\end{aligned}$$

- MLE solves

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\beta}).$$

- NR method can be applied!

Logistic Regression III

- Gradient is

$$\nabla \ell(\boldsymbol{\beta}) := \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i - \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \mathbf{x}_i \right]$$
$$= \sum_{i=1}^{n} \{ y_i - p(\mathbf{x}_i; \boldsymbol{\beta}) \} \mathbf{x}_i$$

- Hessian is

$$\mathbf{H}(\boldsymbol{\beta}) := \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\sum_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\beta}) \{ 1 - p(\mathbf{x}_i; \boldsymbol{\beta}) \} \mathbf{x}_i \mathbf{x}_i^T .$$

- Updating equation:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \mathbf{H}^{-1}(\boldsymbol{\beta}^{(t)}) \nabla \ell(\boldsymbol{\beta}^{(t)}). \quad (1)$$

Generalized Linear Model I

- LR belongs to a more general class of statistical model called generalized linear model (GLM).
- GLM assumes the exponential dispersion family whose density is

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \tag{2}$$

- Given $(y_i,\cdots,y_n)$, the log-likelihood is

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

Generalized Linear Model II

- If $Y \sim f(y;\theta)$, it can be shown that

$$E\left\{\frac{\partial}{\partial\theta}\log f(Y;\theta)\right\} = 0 \tag{3}$$

and

$$-E\left\{\frac{\partial^2}{\partial\theta^2}\log f(Y;\theta)\right\} = E\left[\left\{\frac{\partial}{\partial\theta}\log f(Y;\theta)\right\}^2\right] \tag{4}$$

- For GLM density, (3) yields

$$\mu_i = E(Y_i) = b'(\theta_i),$$

and (4) yields

$$\mathrm{Var}(Y_i) = b''(\theta_i)a(\phi).$$

Generalized Linear Model III

- GLM links $\eta_i = \boldsymbol{\beta}^T \mathbf{x}_i$ to $\mu_i = E(Y_i)$ by a link function $g(\cdot)$, i.e.,

$$\eta_i = g(\mu_i) = \boldsymbol{\beta}^T \mathbf{x}_i, \qquad i = 1, \cdots, n.$$

- We call $g(\cdot)$ the canonical link if $g(\mu_i) = \theta_i$ under (2).
- Notice that $\mu_i = b'(\theta_i)$, and hence $g = (b')^{-1}$ is the canonical link.

Generalized Linear Model IV

- Suppose $n_i y_i \overset{iid}{\sim} Binomial(n_i, p_i)$ (i.e., y_i is relative frequency) then

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

where

$$
\theta_i = \log\left\{ \frac{\pi_i}{1 - \pi_i} \right\}.
$$

- Notice that

$$
b(\theta_i) = \log\{1 + \exp(\theta_i)\}, \quad \text{and} \quad a(\phi) = 1/n_i
$$

Generalized Linear Model V

- Thus we have

$$E(Y_i) = b'(\theta_i) = \frac{\partial}{\partial \theta_i} \log\{1 + \exp(\theta_i)\} = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \pi_i$$

  and

$$\mathrm{Var}(Y_i) = b''(\theta_i) = \frac{\exp(\theta_i)}{\{1 + \exp(\theta_i)\}^2 n_i} = \pi_i(1 - \pi_i)/n_i$$

- Finally, the logistic regression assumes

$$\theta_i = (b')^{-1}(\mu_i) = \log \frac{\pi_i}{1 - \pi_i} = \boldsymbol{\beta}^T \mathbf{x}_i.$$

Generalized Linear Model VI

- The likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} \right\} + \sum_{i=1}^{n} c(y_i, \phi)$$

- The likelihood equation is

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

- By chain rule, we have

$$\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}$$

Generalized Linear Model VII

- Notice that

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)},$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\mathrm{Var}(Y_i)}{a(\phi)},$$

and

$$\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \mathbf{x}_i.$$

- Likelihood equation becomes

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \mathbf{x}_i \times \frac{\partial \mu_i}{\partial \eta_i} = \mathbf{0},$$

where $\frac{\partial \mu_i}{\partial \eta_i}$ depends on the link function used since

$$\mu_i = g^{-1}(\eta_i).$$

Generalized Linear Model VIII

- (Logistic Regression) Suppose $(n_i y_i) \overset{iid}{\sim} Binomial(n_i, \pi_i)$ with $g(\pi_i) = \log \pi_i/(1-\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta} (= \eta_i)$.
- We have

$$\frac{\partial \eta_i}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \log\left(\frac{\pi_i}{1-\pi_i}\right) = \frac{1-\pi_i}{\pi_i} \frac{(1-\pi_i)+\pi_i}{(1-\pi_i)^2} = \frac{1}{\pi_i(1-\pi_i)}$$

The likelihood equation is

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (n_i y_i - n_i \pi_i) \cdot \mathbf{x}_i$$

where

$$\pi_i(\beta) = \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}}.$$

Generalized Linear Model IX

- (Poisson Regression) Suppose $y_i \overset{iid}{\sim} Poisson(\mu_i)$.
- It can be shown that the canonical link of Poisson distribution is $g(\mu_i) = \log(\mu_i)$.
- Poisson regression assumes

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

- We have

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial}{\partial \mu_i} \log \mu_i = \frac{1}{\mu_i}$$

The likelihood equation is

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mu_i} \mu_i \cdot \mathbf{x}_i = \sum_{i=1}^{n} (y_i - \mu_i) \cdot \mathbf{x}_i = \mathbf{0}.$$

where

$$\mu_i = \exp(\boldsymbol{\beta}^T \mathbf{x}_i).$$

Generalized Linear Model X

- To apply NR method, we need both gradient vector Hessian matrix of $\ell(\boldsymbol{\beta})$.
- Recall that

$$\nabla \ell(\boldsymbol{\beta}) = \begin{cases} \sum_{i=1}^n (n_i y_i - n_i \pi_i) \cdot \mathbf{x}_i & \text{(Logistic)} \\ \sum_{i=1}^n (y_i - \mu_i) \cdot \mathbf{x}_i & \text{(Poisson)} \end{cases} = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu})$$

and thus

$$\mathbf{H}(\boldsymbol{\beta}) = \begin{cases} \sum_{i=1}^n n_i \pi_i (1 - \pi_i) \cdot \mathbf{x}_i \mathbf{x}_i^T & \text{(Logistic)} \\ \sum_{i=1}^n \mu_i \cdot \mathbf{x}_i \mathbf{x}_i^T & \text{(Poisson)} \end{cases} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

Generalized Linear Model XI

- LR updating equation is

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - \left\{\mathbf{H}(\boldsymbol{\beta}^{(t)})\right\}^{-1} \nabla\ell(\boldsymbol{\beta}^{(t)})$$

- Fisher scoring method replaces $\mathbf{H}(\boldsymbol{\beta})$ with its expectation $E\{\mathbf{H}(\boldsymbol{\beta})\}$.
- Notice that

$$\begin{aligned} -E\{\mathbf{H}(\boldsymbol{\beta})\} &= -E\left\{\frac{\partial^2 \log f(Y;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\} \\ &= E\left\{\frac{\partial \log f(Y;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\frac{\partial \log f(Y;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right\} = \mathbf{I}(\boldsymbol{\beta}) \end{aligned}$$

  is called information matrix and a very important quantity related to asymptotic variance of $\hat{\boldsymbol{\beta}}$ (MLE).
- For a canonical link, $\mathbf{H}(\boldsymbol{\beta}) = E\{\mathbf{H}(\boldsymbol{\beta})\}$.

Generalized Linear Model XII

- Using matrix notation, (1) becomes

$$
\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + (\mathbf{X}^T\mathbf{W}_{(t)}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}_{(t)}) \\
&= (\mathbf{X}^T\mathbf{W}_{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{(t)}\left\{\mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{W}_{(t)}^{-1}(\mathbf{y}-\boldsymbol{\mu}_{(t)})\right\} \\
&= (\mathbf{X}^T\mathbf{W}_{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{(t)}\mathbf{z}_{(t)}
\end{aligned}
$$

- Let $\tilde{\mathbf{X}} = \mathbf{W}_{(t)}^{1/2}\mathbf{X}$ and $\tilde{\mathbf{y}} = \mathbf{W}_{(t)}^{1/2}\mathbf{z}_{(t)}$
- Then we the updating equation is equivalent to solve the following linear equations:

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \tilde{\mathbf{X}}^T\tilde{\mathbf{y}}$$

which can be easily solved by QR decomposition (Chapter 3)

- This is known as the Iteratively Re-weighted Least Square (IRLS) method.

Generalized Linear Model XIII

	Normal	Poisson	Binomial	Gamma	Inv Gaussian
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$B(n, \pi)/n$	$G(\mu, v)$	$IG(\mu, \sigma^2)$
Support	$(-\infty, \infty)$	$\{0, 1, \cdots\}$	$\{0, \cdots, n\}/n$	$(0, \infty)$	$(0, \infty)$
$a(\phi)$	$\phi = \sigma^2$	1	$1/m$	v^{-1}	σ^2
$b(\theta)$	$\theta^2/2$	e^θ	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$b'(\theta) = E(Y)$	θ	e^θ	$\frac{e^\theta}{1+e^\theta}$	$-1/\theta$	$(-2\theta)^{-1/2}$
$(b')^{-1}(\mu) = g(\mu)$	μ	$\log(\mu)$	$\log \frac{\mu}{1-\mu}$	μ^{-1}	μ^{-2}
$b''(\theta)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

Table: Summary of some popular GLM models.

Reference

- Agresti, A. (2012). Categorical data analysis, 3rd edition. Wiley. Chapter 4.