STATS 205: Hierarchical Linear Models	**Fall 2018**

Lecture 8: Generalized Linear Models

Lecturer: Prof. Jingyi Jessica Li — *Subscriber: Lucia Jimenez and Heather Zhou*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

8.1 GLM Theory

Random structure:

$$Y_i \overset{ind}{\sim} \text{Exponential Family}(\theta_i, \phi)$$
$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

Recall from the last lecture that

$$\mu_i \triangleq \mathbb{E}[Y_i] = b'(\theta_i)$$

and

$$\text{Var}[Y_i] = b''(\theta_i)a_i(\phi)$$

Systematic structure:

$$\eta_i = x_i^T\beta, \text{ “linear predictor”}$$
$$\eta_i = g(\mu_i), \text{ “link function”}$$

8.1.1 Canonical Link Function

Set $\eta_i = \theta_i$. Therefore,

$$\mu_i = b'(\theta_i) = b'(\eta_i)$$
$$\eta_i = (b')^{-1}(\mu_i)$$

and the canonical link function is

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

With a canonical link function $g(\cdot)$, the log likelihood of β becomes

$$\begin{aligned} l(\beta) &\overset{ind}{=} \sum_{i=1}^n \log f(Y_i) \\ &\overset{canonical}{=} \sum_{i=1}^n \frac{y_i\eta_i - b(\eta_i)}{a_i(\phi)} + c(y_i, \phi) \\ &\overset{\eta_i = x_i^T\beta}{=} \sum_{i=1}^n \frac{y_i x_i^T\beta - b(x_i^T\beta)}{a_i(\phi)} + c(y_i, \phi) \end{aligned}$$

Question: what is $\underset{\beta}{\text{argmax}}\ l(\beta)$?

8.2 Review: Optimization Methods

Problem: $f : \mathbb{R}^p \to \mathbb{R}$. Find x^* such that $f(x^*) \geq f(x)$ for all $x \in \mathbb{R}^p$.

8.2.1 Newton-Raphson Method (the most general method)

Notation

Gradient: $\nabla f(x)$ is a p-dimensional vector, the same dimension as x, and

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_p}(x)\right)^T$$

Hessian: $H_f(x)$ is a $p \times p$ matrix, and

$$[H_f(x)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

- If f is strictly concave, $H_f(x)$ is negative definite
- If f is smooth (has derivatives on both sides), $H_f(x)$ is symmetric

N-R for multidimensional optimization

Using 1st order Taylor expansion to approximate $\nabla f(x)$, we have

$$\nabla f(x) \approx \nabla f(x_0) + H_f(x_0)(x - x_0)$$

If x^* is a local maxima, then

$$\nabla f(x_0) + H_f(x_0)(x^* - x_0) \approx \nabla f(x) = 0$$

Therefore,

$$x^* = x_0 - (H_f(x_0))^{-1} \nabla f(x_0)$$

N-R algorithm (iterative)

- Start at an arbitrary value $x^{(0)}$.
- At the k^{th} iteration, $x^{(k+1)} = x^{(k)} - H_f(x^{(k)})^{-1} \nabla f(x^{(k)})$.
- Repeat the above step until convergence.

8.2.2 Fisher Scoring Method

- Fisher scoring figures out a very nice form of Hessian of log likelihood.
- It takes expectation of first and second derivatives since the likelihood is random due to its property of including the samples and responses in its formula.

- $l(\beta; Y_1, Y_2, \cdots, Y_n) \overset{ind}{=} \sum_{i=1}^{n} \log f(Y_i|\beta)$ is the log likelihood to be maximized.

- In the following calculations, we take derivatives of likelihood functions with respect to β.

Gradient (score function)

$$S(\beta) = \nabla l(\beta) = \sum_{i=1}^{n} \frac{\nabla f(Y_i; \beta)}{f(Y_i; \beta)}$$

Note that the score function $S(\beta)$ is a $p \times 1$ random vector.

Property: $\mathbb{E}[S(\beta)] = 0$

Proof:

$$\begin{aligned}
\mathbb{E}[S(\beta)] &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\nabla f(Y_i; \beta)}{f(Y_i; \beta)}\right] \\
&= \sum_{i=1}^{n} \int \frac{\nabla f(y_i; \beta)}{f(y_i; \beta)} f(y_i; \beta) dy_i \\
&\overset{\nabla \leftrightarrow \int}{=} \sum_{i=1}^{n} \nabla \int f(y_i; \beta) dy_i \\
&= \sum_{i=1}^{n} \nabla 1 \\
&= 0
\end{aligned}$$

Hessian $H_l(\beta)$

$$H_l(\beta) = \sum_{i=1}^{n} \frac{f(Y_i; \beta) H_f(\beta) - \nabla f(Y_i; \beta) \cdot (\nabla f(Y_i; \beta))^T}{f^2(Y_i; \beta)}$$

Fisher proposed to simplify $H_l(\beta)$ as $\mathbb{E}[H_l(\beta)]$.

Notice that $\forall\ i$

$$\begin{aligned}
\mathbb{E}\left[\frac{f(Y_i; \beta) H_f(\beta)}{f^2(Y_i; \beta)}\right] &= \int \frac{f(y_i; \beta) H_f(\beta)}{f^2(y_i; \beta)} f(y_i; \beta) dy_i \\
&= \int H_f(\beta) dy_i \\
&= \nabla^2 1 \\
&= 0
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[H_l(\beta)] &= -\mathbb{E}\left[\sum_{i=1}^{n} \frac{\nabla f(Y_i;\beta)\cdot(\nabla f(Y_i;\beta))^T}{f^2(Y_i;\beta)}\right] \\
&= -\sum_{i=1}^{n} \mathbb{E}\left[S_i(\beta)\cdot(S_i(\beta))^T\right] && \text{where } S_i(\beta) = \frac{\nabla f(Y_i;\beta)}{f(Y_i;\beta)} \\
&= -\sum_{i=1}^{n} \mathrm{Cov}(S_i(\beta)) && \text{since } \mathbb{E}[S_i(\beta)] = 0 \\
&= -I(\beta)
\end{aligned}$$

where $I(\beta)$ is Fisher's information.

Note: $(I(\beta))^{-1}$ is the asymptotic covariance matrix of $\hat{\beta}_{MLE}$.

Optimization algorithm

- NR update: $\beta^{(k+1)} = \beta^{(k)} - H_l(\beta^{(k)})^{-1}S(\beta^{(k)})$
- FS update: $\beta^{(k+1)} = \beta^{(k)} + I(\beta^{(k)})^{-1}S(\beta^{(k)})$

Example: logistic regression

$$\begin{aligned}
l(\beta) &= \sum_{i=1}^{n}\left[Y_i \log(\frac{\pi_i}{1-\pi_i}) + \log(1-\pi_i)\right] \\
&\overset{\log\frac{\pi_i}{1-\pi_i}=x_i^T\beta}{=\!=} \sum_{i=1}^{n}\left[Y_i x_i^T\beta + \log(1+e^{x_i^T\beta})\right]
\end{aligned}$$

Therefore,

$$\begin{aligned}
S(\beta) = \nabla l(\beta) &= \sum_{i=1}^{n}\left[Y_i x_i - \frac{e^{x_i^T\beta}x_i}{1+e^{x_i^T\beta}}\right] \\
H_l(\beta) &= -\sum_{i=1}^{n}\left[\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})^2}x_i\cdot x_i^T\right]
\end{aligned}$$

Thus,

$$\mathbb{E}[H_l(\beta)] = H_l(\beta)$$

That is, Fisher's Scoring Method is identical to Newton-Raphson Method in this scenario.

8.2.3 Iteratively Reweighted Least Squares (IRLS) Based on FS

$$l(\beta) = \sum_{i=1}^{n}\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i,\phi) \triangleq \sum_{i=1}^{n} l_i$$

To get $\nabla l(\beta)$, we look at ∇l_i first and we have

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

Notice:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{Y_i - b'(\theta_i)}{a_i(\phi)} = \frac{Y_i - \mu_i}{a_i(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} \text{ since } \mu_i = b'(\theta_i), \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \text{ since } \eta_i = x_i^T \beta = \sum_{j=1}^{p} x_{ij}\beta_j$$

Therefore,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{Y_i - \mu_i}{a_i(\phi)} \frac{1}{b''(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

Now, define

$$w_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{a_i(\phi) b''(\theta_i)}$$

and we have

$$\frac{\partial l_i}{\partial \beta_j} = (Y_i - \mu_i) w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}$$

$$\frac{\partial l}{\partial \beta_j} = \sum \frac{\partial l_i}{\partial \beta_j}$$

Therefore,

$$S(\beta) = \nabla l(\beta) = X^T W (Y - \mu) \frac{d\eta}{d\mu}$$

where

$$W = \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix}_{n \times n}$$

and

$$(Y - \mu)\frac{d\eta}{d\mu} = \begin{bmatrix} (Y_1 - \mu_1)\frac{d\eta_1}{d\mu_1} \\ \vdots \\ (Y_n - \mu_n)\frac{d\eta_n}{d\mu_n} \end{bmatrix}$$

To get $H_l(\beta)$, we have

$$\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = (Y_i - \mu_i) \frac{\partial (w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij})}{\partial \beta_k} + \frac{\partial (Y_i - \mu_i)}{\partial \beta_k} w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}$$

$$\begin{aligned} \mathbb{E}\left[\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}\right] &= \mathbb{E}\left[(Y_i - \mu_i) \frac{\partial (w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij})}{\partial \beta_k}\right] + \mathbb{E}\left[\frac{\partial (Y_i - \mu_i)}{\partial \beta_k} w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}\right] \\ &= 0 - \mathbb{E}\left[\frac{\partial \mu_i}{\partial \beta_k} w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}\right] \\ &= -\mathbb{E}\left[\frac{\partial \eta_i}{\partial \beta_k} w_i x_{ij}\right] \\ &= -x_{ik} w_i x_{ij} \end{aligned}$$

So

$$I(\beta) = -\mathbb{E}[H_l(\beta)] = \underset{(p\times n)}{X^T}\underset{(n\times n)}{W}\underset{(n\times p)}{X}$$

Therefore, by FS,

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} + (I(\beta^{(k)}))^{-1}S(\beta^{(k)})\\ &= \beta^{(k)} + (X^TW^{(k)}X)^{-1}X^TW^{(k)}(Y-\mu^{(k)})\frac{d\eta^{(k)}}{d\mu^{(k)}}\end{aligned}$$