Lecture 5. Generalized Linear Models (cont.)

Nan Ye

School of Mathematics and Physics
University of Queensland

This Lecture

- Fisher scoring for GLM
- Properties of MLE
- GLM with canonical link

Fisher Scoring for GLM

Recall: Fisher scoring

- An general algorithm for finding an MLE.
- Start with some $\beta^{(0)}$. At iteration $t \geq 0$,

$$\beta^{(t+1)} = \beta^{(t)} + I^{-1}(\beta^{(t)}) \nabla \ell(\beta^{(t)}).$$

where $I(\beta) = -\mathbb{E} \nabla^2 \ell(\beta)$ (known as *Fisher information*).

Log-likelihood for GLM

- Given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, our objective is to maximize the log-likelihood

$$\ell(\beta) = \sum_i \ln p(y_i \mid \mathbf{x}_i, \beta).$$

- Recall: $p(y \mid \mathbf{x}, \beta)$ can be explicitly computed as

$$p(y \mid \mathbf{x}, \beta) = \exp\left(\frac{\eta y - A(\eta)}{b(\phi)} + c(y, \phi)\right),$$

where $\eta = A'^{-1}(g^{-1}(\beta^\top \mathbf{x}))$.

Fisher scoring for GLM

- Let $\mu_i = \mathbb{E}(Y_i \mid \mathbf{x}_i, \beta) = g(\mathbf{x}_i^\top \beta)$ and $V_i = \mathrm{var}(Y_i \mid \mathbf{x}_i, \beta)$.
- The gradient, or score function, is

$$\nabla \ell(\beta) = \sum_i \frac{y_i - \mu_i}{g'(\mu_i) V_i} \mathbf{x}_i.$$

- The Fisher information is

$$I(\beta) = -\sum_i \frac{1}{g'(\mu_i)^2 V_i} \mathbf{x}_i \mathbf{x}_i^\top.$$

No specific parametrization of the exponential family is required. Choose whichever is more convenient for computing the variances.

Interpretation

- Gradient is a linear combination of input $\mathbf{x}_i$'s.

  Weight of $\mathbf{x}_i$ is

  - proportional to $y_i - \mu_i$ (mean's quality as a predictor),
  - inversely proportional to V_i (variance of the response),
  - proportional to $\frac{1}{g'(\mu_i)} = \frac{d\mu_i}{d(\mathbf{x}_i^\top \beta)}$ (rate of change of mean in the linear predictor).

- Fisher information is a linear combination of $\mathbf{x}_i\mathbf{x}_i^\top$'s.

  Weight of $\mathbf{x}_i\mathbf{x}_i$ is

  - inversely proportional to V_i,
  - proportional to $\frac{1}{g'(\mu_i)^2}$.

Example 1. Ordinary least squares

- Recall: $Y_i \overset{ind}{\sim} N(\mathbf{x}_i^\top \beta, \sigma^2)$.
- We have $\mu_i = \mathbf{x}_i^\top \beta$, $V_i = \sigma^2$, $g(\mu) = \mu$, $g'(\mu) = 1$, thus

$$\nabla \ell(\beta) = \sum_i \frac{y_i - \mathbf{x}_i^\top \beta}{\sigma^2} \mathbf{x}_i = \frac{1}{\sigma^2}(\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\beta),$$

$$I(\beta) = -\sum \frac{1}{\sigma^2} \mathbf{x}_i \mathbf{x}_i^\top = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X},$$

where $\mathbf{X}$ is the design matrix.

- For any $\beta^{(0)}$, we have

$$
\begin{aligned}
\beta^{(1)} &= \beta^{(0)} + \left(-\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\frac{1}{\sigma^2}(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\beta^{(0)})\right) \\
&= \beta^{(0)} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - \beta^{(0)} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}
$$

- This is exactly the MLE that we are familiar with.
- Thus the MLE is found after one Fisher scoring iteration.

Derivation

- It suffices to work out the case with one example $(\mathbf{x}, y)$,

$$\ell(\beta) = \ln p(y \mid \mathbf{x}, \beta),$$

and then applying a summation over the examples to obtain the general case.

- For the gradient, using the chain rule,

$$\nabla \ell(\beta) = \frac{d\ell}{d\eta} \nabla \eta(\beta) = \frac{y-\mu}{b(\phi)} \nabla \eta(\beta)$$

To find $\nabla \eta(\beta)$, differentiate $g(A'(\eta)) = g(\mu) = \mathbf{x}^\top \beta$

$$g'(A'(\eta))A''(\eta) \nabla \eta(\beta) = \mathbf{x}.$$

Hence we have $\nabla \eta(\beta) = \frac{1}{g'(\mu)A''(\eta)}\mathbf{x}$, and thus

$$\nabla \ell(\beta) = \frac{d\ell}{d\eta} \nabla \eta(\beta) = \frac{y-\mu}{b(\phi)} \frac{1}{g'(\mu)A''(\eta)}\mathbf{x} = \frac{y-\mu}{g'(\mu)V},$$

where $V = \mathrm{var}(Y \mid \mathbf{x}, \beta) = b(\phi)A''(\eta)$.

- For Fisher information, differentiate $\nabla \ell(\beta)$ using the product rule

$$\nabla^2 \ell(\beta) = \frac{1}{g'(\mu)A''(\eta)}\mathbf{x}\,\nabla^\top \left(\frac{y-\mu}{b(\phi)}\right) + \frac{y-\mu}{b(\phi)}\,\nabla^\top \left(\frac{1}{g'(\mu)A''(\eta)}\mathbf{x}\right)$$

Using $\nabla(y-\mu) = \nabla \mu$ and $\mathbb{E}(y-\mu) = 0$, we have

$$I(\beta) = \mathbb{E}(-\nabla^2 \ell(\beta)) = \frac{1}{g'(\mu)b(\phi)A''(\eta)}\mathbf{x}\,\nabla^\top \mu(\beta).$$

To find $\nabla \mu(\beta)$, differentiate $g(\mu) = \mathbf{x}^\top \beta$

$$g'(\mu)\,\nabla \mu(\beta) = \mathbf{x}.$$

Hence $\nabla \mu(\beta) = \frac{1}{g'(\mu)}\mathbf{x}$, thus

$$I(\beta) = \frac{1}{g'(\mu)^2 b(\phi)A''(\eta)}\mathbf{x}\mathbf{x}^\top = \frac{1}{g'(\mu)^2 V}\mathbf{x}\mathbf{x}^\top.$$

Matrix form

- Let $\mathbf{y} = (y_1, \ldots, y_n)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, $\mathbf{X}$ be the design matrix,

$$\mathbf{W} = \mathrm{diag}\left(\frac{1}{g'(\mu_1)^2 V_1}, \ldots, \frac{1}{g'(\mu_n) V_n}\right),$$
$$\mathbf{G} = \mathrm{diag}(g'(\mu_1), \ldots, g'(\mu_n)).$$

- Then we have

$$\nabla \ell(\beta) = \mathbf{X}^\top \mathbf{W}(\mathbf{G}\mathbf{y} - \mathbf{G}\boldsymbol{\mu}),$$
$$I(\beta) = \mathbf{X}^\top \mathbf{W}\mathbf{X}.$$

- Thus Fisher scoring updates β to β'

$$\begin{aligned}\beta' &= \beta + (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}(\mathbf{G}\mathbf{y} - \mathbf{G}\boldsymbol{\mu}) \\ &= (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}(\mathbf{G}\mathbf{y} - \mathbf{G}\boldsymbol{\mu} + \mathbf{X}\beta).\end{aligned}$$

Fisher scoring as IRLS

- Let $\mathbf{z} = \mathbf{G}\mathbf{y} - \mathbf{G}\boldsymbol{\mu} + \mathbf{X}\beta$, then Fisher scoring update is

$$\beta' = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{z},$$

- β' is the solution of the weighted least squares problem

$$\min_{\tilde{\beta}} (\mathbf{z} - \mathbf{X}\tilde{\beta})\mathbf{W}(\mathbf{z} - \mathbf{X}\tilde{\beta}).$$

- Fisher scoring is thus an instance of iteratively reweighted least squares (IRLS) algorithm.

Properties of MLE

Assumption

The model is well-specified, that is, each y_i is independently drawn from $p(Y \mid \mathbf{x}_i, \beta^*)$, that is, the GLM with parameter β^*.

Asymptotic normality

Under appropriate regularity conditions, the MLE $\hat{\beta}$ is asymptotically normally distributed with mean β^*, and covariance $I^{-1}(\beta^*)$.

$I(\beta)$ is linear in n, thus the entries of the covariance matrix is of the order $1/n$.

Confidence interval

Marginal confidence interval for β_i is given by

$$\hat{\beta}_i \pm z_{\alpha/2}\sigma_i,$$

where $\sigma_i = \sqrt{I^{-1}(\beta^*)_{ii}}$. This is approximated by

$$\hat{\beta}_i \pm z_{\alpha/2}\hat{\sigma}_i,$$

where $\hat{\sigma}_i = \sqrt{I^{-1}(\hat{\beta})_{ii}}$.

Testing significance of effect

- We want to test whether the i-th covariate has a significant effect

$$H_0 \quad \beta_i^* = 0, \qquad\qquad H_1 \quad \beta_i^* \neq 0.$$

- Under H_0, the Wald statistic $T = \frac{\hat{\beta}_i}{\hat{\sigma}_i}$ is asymptotically standard normal

$$T \sim N(0,1).$$

- At significance level α, reject H_0 iff $|T| \geq z_{\alpha/2}$.

Remark

- With a mis-specified model, asymptotic normality still holds, but the mean and the covariance matrix of the asymptotic distribution now depend on both the model class and the *unknown* true distribution.
- The confidence interval and the distribution of Wald's statistics cannot be computed, and can only be applied (*with caution*) if the model is not too much away from reality.

GLM with Canonical Link

Motivation

- For OLS and logistic regression, both have the linear predictor $\mathbf{x}^\top \beta$ as the natural parameter.
- GLMs with this property are mathematically appealing to work with.

Canonical link

- A link function $g(\cdot)$ is called a canonical link if $g(\mu) = \eta$, that is, $\eta = \beta^\top \mathbf{x}$.
- For a natural exponential family, the canonical link is A'^{-1}.
- A GLM using a canonical link can be written down as

$$p(y \mid \mathbf{x}, \beta) = \exp\left(\frac{y\mathbf{x}^\top \beta - A(\mathbf{x}^\top \beta)}{b(\phi)} + c(y, \phi) \right),$$

where A is from the natural form of the exponential family.

Examples

Exponential family	Canonical link	GLM
Normal	$g(\mu) = \mu$	OLS
Poisson	$g(\mu) = \ln \mu$	Poisson regression
Binomial	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	Logistic regression
Gamma	$g(\mu) = \mu^{-1}$	

Remark

- The form of GLM with canonical link is mathematically convenient.
- However, it does not imply that canonical link necessarily leads to a better model.

What You Need to Know

- Fisher scoring for GLMs

  update rule, interpretation, example, derivation, matrix form, IRLS
- Properties of MLE

  when model is well-specified, and when model is mis-specified
- Models with canonical links

  mathematically convenient, but not necessarily a better model.