Lec 4: Generalized Linear Model

Yanjun Han

Sept. 26, 2023

Generalized linear model.

Setting. For $i = 1, 2, \cdots, n$, let $y_i \overset{ind}{\sim} p_{\theta_i}(y_i) = \exp(\langle \theta_i, T(y_i)\rangle - A(\theta_i))h(y_i)$.

where $\theta_i = (\langle x_i, \beta_1\rangle, \langle x_i, \beta_2\rangle, \cdots, \langle x_i, \beta_d\rangle) \in \mathbb{R}^d$

- $x_i \in \mathbb{R}^p$: feature / covariate
- $(\beta_1, \cdots, \beta_d) \in \mathbb{R}^{p\times d}$: regression coefficients
- written in matrix form : $\theta_i = \beta^T x_i$

MLE.

$$\begin{aligned}
\hat{\beta} &= \arg\max_\beta \prod_{i=1}^n p_{\theta_i}(y_i) \\
&= \arg\max_\beta \sum_{i=1}^n \left(\langle \beta^T x_i, T(y_i)\rangle - A(\beta^T x_i)\right) \\
&= \arg\max_\beta \underbrace{\mathrm{Tr}\left(\sum_{i=1}^n T(y_i) x_i^T \cdot \beta\right)}_{\text{linear in } \beta} - \underbrace{\sum_{i=1}^n A(\beta^T x_i)}_{\text{convex in } \beta}
\end{aligned}$$

Estimating equation $(d=1)$: $\sum_{i=1}^n T(y_i) x_i = \sum_{i=1}^n A'(\hat{\beta}^T x_i) x_i$.

The computation of MLE is a convex problem, thus efficient.

In R: model <- glm(y ~ X , family).

Examples. 1. Linear regression.

$$y_i \sim N(\theta_i, 1) = N(\beta^T x_i, 1)$$

$$\Rightarrow \hat{\beta} = \arg\min_\beta \sum_{i=1}^n (y_i - \beta^T x_i)^2 = \arg\min_\beta \|y - \underset{\mathbb{R}^{n\times p}}{X}\beta\|_2^2$$

2. Logistic regression.

$$y_i \sim \mathrm{Bern}\left(\frac{1}{1+e^{-\theta_i}}\right) = \mathrm{Bern}\left(\frac{1}{1+e^{-\beta^T x_i}}\right)$$

$$\begin{aligned}
\Rightarrow \hat{\beta} &= \arg\max_\beta \sum_{i=1}^n \left(y_i \log \frac{1}{1+e^{-\beta^T x_i}} + (1-y_i)\log \frac{e^{-\beta^T x_i}}{1+e^{-\beta^T x_i}}\right) \\
&= \arg\max_\beta \sum_{i=1}^n \left(y_i \beta^T x_i - \log(1+e^{\beta^T x_i})\right)
\end{aligned}$$

2'. Probit model.

$$Y_i \sim \mathrm{Bern}(\Phi(\theta_i)) = \mathrm{Bern}(\Phi(\beta^T x_i)).$$

where Φ is the standard normal CDF:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

MLE: $$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \left(y_i \log \Phi(\beta^T x_i) + (1-y_i) \log (1-\Phi(\beta^T x_i)) \right)$$

Lemma. The above objective is concave in β.

Pf. For $f(x) = \log \Phi(x)$:

$$f'(x) = \frac{\varphi(x)}{\Phi(x)}, \quad f''(x) = \frac{\varphi' \Phi - \varphi^2}{\Phi^2} = -\frac{(x\Phi + \varphi)\varphi}{\Phi^2}.$$

Gaussian Mills ratio:

$$1 - \Phi(x) < \frac{\varphi(x)}{x}, \quad x > 0$$

$$\Rightarrow x\Phi(x) + \varphi(x) > 0, \; x < 0 \Rightarrow f''(x) < 0.$$

> In an exponential family, there could be more than one parametrizations such that the MLE computation in the corresponding GLM is a convex problem.

3. Poisson regression.

$$Y_i \sim \mathrm{Poi}(e^{\theta_i}) = \mathrm{Poi}(e^{\beta^T x_i})$$

$$\Rightarrow \hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \left(T(Y_i)\beta^T x_i - A(\beta^T x_i) \right)$$

$$= \arg\max_{\beta} \sum_{i=1}^{n} \left(Y_i \beta^T x_i - e^{\beta^T x_i} \right).$$

4. Multinomial logit regression.

Recall that $\theta = (\theta_1, \cdots, \theta_k)$

$$T(y) = (1(y=1), 1(y=2), \cdots, 1(y=k))$$

$$A(\theta) = \log(e^{\theta_1} + \cdots + e^{\theta_k})$$

Model: $\mathbb{P}(y_i = j \mid x_i) = \dfrac{e^{\beta_j^T x_i}}{e^{\beta_1^T x_i} + e^{\beta_2^T x_i} + \cdots + e^{\beta_k^T x_i}}$.

MLE:

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \Big(1(y_i=1)\beta_1^T x_i + 1(y_i=2)\beta_2^T x_i + \cdots + 1(y_i=k)\beta_k^T x_i - \log\Big(\sum_{j=1}^{k} e^{\beta_j^T x_i}\Big)\Big)$$

$$= \arg\max_{\beta} \sum_{j=1}^{k} \beta_j^T \sum_{i: y_i=j} x_i - n\log\Big(\sum_{j=1}^{k} e^{\beta_j^T x_i}\Big).$$

Note: the MLE is not unique, as $(\beta_1, \cdots, \beta_k)$ and $(\beta_1 + c, \cdots, \beta_k + c)$ give the same objective.

So we can assume that $\beta_1 = 0$.

4' Ordered logit model (ordinal regression).

Suppose y_i could take k values with ordered relationship.

Model:

$$\log \frac{\mathbb{P}(y_i \le j)}{\mathbb{P}(y_i > j)} = \alpha_j + \beta^T x \quad (j = 1, 2, \cdots, k-1)$$

or equivalently,

$$\mathbb{P}(y_i \le j) = \frac{1}{1 + e^{-(\alpha_j + \beta^T x)}}.$$

Proportional odds assumption: the difference in the log-odds

$$\log \frac{\mathbb{P}(y_i \le j+1)}{\mathbb{P}(y_i > j+1)} - \log \frac{\mathbb{P}(y_i \le j)}{\mathbb{P}(y_i > j)}$$

is independent of x. More on this in Lecture 5.

MLE:

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{(\alpha,\beta)} \sum_{i=1}^{n} \Big(\sum_{j=1}^{k} 1(y_i = j) \log \mathbb{P}(y_i = j)\Big)$$

$$= \arg\max_{(\alpha,\beta)} \sum_{i=1}^{n} \Big(\sum_{j=1}^{k} 1(y_i = j) \times \log\Big(\frac{1}{1+e^{-(\alpha_j + \beta^T x)}} - \frac{1}{1+e^{-(\alpha_{j-1} + \beta^T x)}}\Big)\Big)$$

where $\alpha_0 \triangleq 0$, $\alpha_k \triangleq +\infty$.

$$\boxed{\text{Exercise (HW): show that the log-likelihood is concave in } (\alpha, \beta).}$$

Variance of MLE.

In the sequel we assume that $d=1$ for simplicity, i.e. $\beta \in \mathbb{R}^1$.

$$\text{F.O.C. for MLE: } 0 = \sum_{i=1}^{n} \left(T(y_i) - A'(x_i^T \hat{\beta}^{MLE}) \right) x_i$$

$$= \sum_{i=1}^{n} \left(A'(x_i^T \beta) - A'(x_i^T \hat{\beta}^{MLE}) \right) x_i$$

$$+ \underbrace{\sum_{i=1}^{n} \left(T(y_i) - A'(x_i^T \beta) \right) x_i}_{\mathrm{Cov}(\cdot) = \sum_{i=1}^{n} A''(x_i^T\beta) x_i x_i^T}$$

Delta method (Taylor expansion):

$$\text{first term} \approx \left(\sum_{i=1}^{n} A''(x_i^T \beta) x_i x_i^T \right) (\beta - \hat{\beta}^{MLE})$$

$$\boxed{\mathrm{Cov}_\beta(\hat{\beta}^{MLE}) \approx \left(\sum_{i=1}^{n} A''(x_i^T \beta) x_i x_i^T \right)^{-1}.}$$

Fisher information.

Def. For a (regular) class of probability distributions $(p_\theta)_{\theta \in \mathbb{R}^d}$, the Fisher information at $\theta = \theta_0$ is defined as

$$I(\theta_0) = \mathbb{E}_{\theta_0} \left[- \frac{\partial^2 \log p_\theta(y)}{\partial \theta^2} \Big|_{\theta = \theta_0} \right]$$

Side note: $\dot{\ell}_{\theta_0}(y) = \dfrac{\partial \log p_\theta(y)}{\partial \theta} \Big|_{\theta=\theta_0}$ (score)

$\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0}(y)] = 0$

$\mathrm{Cov}_{\theta_0}(\dot{\ell}_{\theta_0}(y)) = I(\theta_0)$

In GLM: $\ell_\beta(x,y) = \sum_{i=1}^{n} \log p_{\theta_i}(y_i) = \sum_{i=1}^{n} \left(T(y_i)\beta^T x_i - A(\beta^T x_i)\right)$
$+ \text{const}(x,y)$

$$\dot{\ell}_\beta(x,y) = \frac{\partial}{\partial\beta}\ell_\beta(x,y) = \sum_{i=1}^{n} \underbrace{\left(T(y_i) - A'(\beta^T x_i)\right)}_{\text{has mean zero}} x_i$$

$$\ddot{\ell}_\beta(x,y) = \frac{\partial}{\partial\beta}\dot{\ell}_\beta(x,y) = -\sum_{i=1}^{n} A''(\beta^T x_i) x_i x_i^T$$

$$\Rightarrow \quad I(\beta) = \mathbb{E}[-\ddot{\ell}_\beta(x,y)] = \sum_{i=1}^{n} A''(\beta^T x_i) x_i x_i^T.$$

(Asymptotic) Cramér-Rao bound: $I(\theta)^{-1}$ is the "best" covariance of any asymptotically unbiased estimator $\hat{\theta}$ for θ as $n \to \infty$.

Asymptotic efficiency of MLE: $\hat{\theta}^{MLE}$ asymptotically achieves the Cramér-Rao bound.

Bootstrap estimate for $\text{Cov}(\hat{\beta}^{MLE})$: same as Lecture 3.

Inference in GLM.

Deviance. Deviance for data point i is

$$D_i(\hat{\beta};\beta) = D(x_i^T\hat{\beta}; x_i^T\beta) = 2\left(A(x_i^T\beta) - A(x_i^T\hat{\beta}) - A'(x_i^T\hat{\beta})x_i^T(\beta - \hat{\beta})\right).$$

(a generalization of "training error" $(x_i^T\hat{\beta} - x_i^T\beta)^2$ in linear regression)

Total deviance. $D_+(\hat{\beta};\beta) = \sum_{i=1}^{n} D_i(\hat{\beta};\beta)$.

Hoeffding's formula.

$$D_+(\hat{\beta}^{MLE};\beta) = 2\log\frac{p_{\hat{\beta}^{MLE}}(y^n | x^n)}{p_\beta(y^n | x^n)} \quad \text{(Pf: see HW)}$$

Deviance table.

Setting: $\underset{p\times 1}{\beta} = (\underset{p^{(1)}\times 1}{\beta^{(1)}}, \underset{p^{(2)}\times 1}{\beta^{(2)}}, \cdots, \underset{p^{(J)}\times 1}{\beta^{(J)}})$ with $\sum_{j=1}^{J} p^{(j)} = p$.

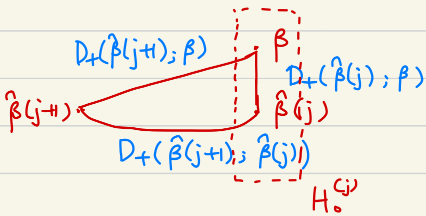
Target: for each $j = 0, 1, \cdots, J$, test if $\beta^{(j+1)} = \cdots = \beta^{(J)} = 0$. $(H_0^{(j)})$

Notation: let $\hat\beta(j)$ be the MLE assuming that $\hat\beta^{(j+1)} = \cdots = \hat\beta^{(J)} = 0$.

$(\hat\beta(0) = 0 \quad , \quad \hat\beta(J) = \hat\beta^{MLE})$

$$\boxed{\begin{array}{l}\text{Deviance additivity theorem: if } \beta \in H_0^{(j)}, \\ \qquad D_+(\hat\beta(j+1); \hat\beta(j)) = D_+(\hat\beta(j+1); \beta) - D_+(\hat\beta(j); \beta) \\ \text{(Pf: see HW)}\end{array}}$$

A pictorial illustration:



Generalized likelihood-ratio test:

$$D_+(\hat\beta(j+1); \hat\beta(j)) \sim \chi^2_{p^{(j+1)}} \text{ under } H_j : \beta^{(j+1)} = \cdots = \beta^{(J)} = 0.$$

Deviance table.

MLE	$2\times$ log-likelihood	difference	compare with
$\hat\beta(0) = 0$	$2\ell_{\hat\beta(0)}$	$D_+(\hat\beta(1); \hat\beta(0))$	$\chi^2_{p^{(1)}}$
$\hat\beta(1)$	$2\ell_{\hat\beta(1)}$	$D_+(\hat\beta(2); \hat\beta(1))$	$\chi^2_{p^{(2)}}$
$\hat\beta(2)$	$2\ell_{\hat\beta(2)}$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$D_+(\hat\beta(J); \hat\beta(J-1))$	$\chi^2_{p^{(J)}}$
$\hat\beta(J) = \hat\beta^{MLE}$	$2\ell_{\hat\beta(J)}$		

Model selection. (Assuming $p^{(1)} = \cdots = p^{(J)} = 1$)

1. AIC (Akaike information criterion)

$$\boxed{\begin{aligned} j^{AIC} &= \underset{j \in \{0,1,\cdots,J\}}{\operatorname{argmin}} \; -D_+(j) + 2j \\ &= \underset{j \in \{0,1,\cdots,J\}}{\operatorname{argmin}} \; -2\ell_{\hat\beta(j)} + 2j \end{aligned}}$$

2. BIC (Bayesian information criterion).

$$\boxed{\begin{aligned} j^{BIC} &= \underset{j \in \{0,1,\cdots,J\}}{\operatorname{argmin}} \; -D_+(j) + j \ln n \\ &= \underset{j \in \{0,1,\cdots,J\}}{\operatorname{argmin}} \; -2\ell_{\hat\beta(j)} + j \ln n \end{aligned}}$$

3. Lasso.

$$\boxed{\hat\beta^{Lasso} = \underset{\beta}{\operatorname{argmin}} \; -\frac{1}{n} \sum_{i=1}^{n} \log p_{x_i^T\beta}(y_i) + \lambda \|\beta\|_1}$$

- λ is typically chosen by cross validation.

Application: Density estimation via Lindsey's method

Given i.i.d. $z_1, \cdots, z_n \sim p$, aim to fit

$$p \approx p_\theta = \exp(\langle \theta, T(z) \rangle - A(\theta))\, h(z)$$

- known: $T(\cdot)$, $h(\ )$ • unknown: $\theta \in \mathbb{R}^d$.

Problem with MLE: log-partition function $A(\theta)$ untractable (more in Lec 6)

Lindsey's method

- Suppose $\mathcal{Z} \subseteq \mathbb{R}$, and $\mathcal{Z} = \mathcal{Z}_1 \cup \mathcal{Z}_2 \cup \cdots \cup \mathcal{Z}_K$, with
$$\mathcal{Z}_k = [\, z_k - \tfrac{\Delta_k}{2},\ z_k + \tfrac{\Delta_k}{2} \,].$$
- For small Δ_k,
$$\mathbb{P}(z \in \mathcal{Z}_k) = \int_{\mathcal{Z}_k} p_\theta(z)\,dz$$
$$\approx \exp(\langle \theta, T(z_k)\rangle - A(\theta))\, h(z_k)\Delta_k =: p_k.$$
- For $y_k = \#\{z_i \in \mathcal{Z}_k\}$, then
$$(y_1, \cdots, y_K) \sim \text{Multi}(n; (p_1, \cdots, p_K))$$
- Poisson trick: fit
$$y_k \overset{\text{ind.}}{\sim} \text{Poi}(e^{\langle \theta, T(z_k)\rangle + \log(h(z_k)\Delta_k) + \theta_0})$$
This is a Poisson GLM!
- Poisson conditioning property:

$$\boxed{\begin{array}{l} \text{if } y_i \overset{\text{ind.}}{\sim} \text{Poi}(\lambda_i), \text{ then} \\ (y_1, \cdots, y_K) \mid \sum_{k=1}^K y_k = n \sim \text{Multi}\left(n; \left(\frac{\lambda_1}{\sum_k \lambda_k}, \cdots, \frac{\lambda_K}{\sum_k \lambda_k}\right)\right) \end{array}}$$

Therefore, $(y_1, \cdots, y_K) \mid \sum_{k=1}^K y_k = n \sim \text{Multi}(n; (q_1, \cdots, q_K))$, with
$$q_k = \frac{\exp(\langle \theta, T(z_k)\rangle + \log(h(z_k)\Delta_k) + \theta_0)}{\sum_j \exp(\langle \theta, T(z_j)\rangle + \log(h(z_j)\Delta_j) + \theta_0)}$$
$$\propto \exp(\langle \theta, T(z_k)\rangle)\, h(z_k)\, \Delta_k = p_k.$$

- Think: what does θ_0 represent?