Lec 3: Parameter Estimation & Inference

Yanjun Han

Sept 19, 2023

Given i.i.d. $y_1, \dots, y_n \sim p_\theta(y) = \exp(\langle \theta, T(y) \rangle - A(\theta)) h(y)$.

This lecture:

<u>Parameter estimation</u>: estimate θ or functions of θ

<u>Inference</u>: test $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$

Maximum likelihood estimator (MLE)

$$
\begin{aligned}
\hat{\theta}_n &= \arg\max_\theta \prod_{i=1}^n p_\theta(y_i) \\
&= \arg\max_\theta \sum_{i=1}^n \log p_\theta(y_i) \\
&= \arg\max_\theta \underbrace{\langle \theta, \sum_{i=1}^n T(y_i) \rangle - n A(\theta)}_{\text{concave in } \theta}
\end{aligned}
$$

F.O.C.: $\quad 0 = \sum_{i=1}^n T(y_i) - n \nabla A(\hat{\theta}_n)$, or

$$
\boxed{\nabla A(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^n T(y_i)}
$$

- As $\mu_\theta := \mathbb{E}_\theta[T(y)] = \nabla A(\theta)$, the MLE $\hat{\theta}_n$ is chosen so that the "true mean" matches the "sample mean".
- The MLE either admits a closed-form expression, or is the solution to a convex optimization problem.

Example: Poisson family.

Recall that $y \sim \text{Poi}(\lambda)$, $\theta = \log \lambda$, $T(y) = y$, $A(\theta) = e^\theta$.

Therefore.

MLE for θ: $\quad e^{\hat{\theta}_n} = \frac{1}{n} \sum_{i=1}^n y_i \Rightarrow \hat{\theta}_n = \log\left(\frac{1}{n} \sum_{i=1}^n y_i\right)$

MLE for λ: $\quad \hat{\lambda}_n = e^{\hat{\theta}_n} = \frac{1}{n} \sum_{i=1}^n y_i$.

Variance of the MLE.

1. (Exact) variance for $\mu_{\hat{\theta}_n} = \nabla A(\hat{\theta}_n)$:

$$\nabla A(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} T(y_i)$$
$$\Rightarrow \quad \mathrm{Cov}_\theta(\nabla A(\hat{\theta}_n)) = \mathrm{Cov}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} T(y_i)\right)$$
$$\Rightarrow \quad \mathrm{Cov}_\theta(\nabla A(\hat{\theta}_n)) = \frac{1}{n}\nabla^2 A(\theta)$$

In reality we don't know θ, so we typically use

$$\boxed{\mathrm{Cov}_\theta(\nabla A(\hat{\theta}_n)) \approx \frac{1}{n}\nabla^2 A(\hat{\theta}_n)}$$

2. Approximate variance: delta method

Question: Suppose $\hat{\theta}_n \approx \theta$ and $f(\cdot)$ is differentiable at θ.
How is $\mathrm{Var}(f(\hat{\theta}_n))$ related to $\mathrm{Var}(\hat{\theta}_n)$?

Idea of delta method: suppose $|\hat{\theta}_n - \theta| = O_{P_\theta}(r_n)$ with $r_n \to 0$.
Then

$$f(\hat{\theta}_n) = f(\theta) + f'(\theta)(\hat{\theta}_n - \theta) + o_{P_\theta}(r_n)$$
$$\Rightarrow \quad \mathrm{Var}(f(\hat{\theta}_n)) = \mathrm{Var}[f(\theta) + f'(\theta)(\hat{\theta}_n - \theta)] + o_{P_\theta}(r_n^2)$$
$$= f'(\theta)^2 \cdot \mathrm{Var}(\hat{\theta}_n) + o_{P_\theta}(r_n^2)$$

So we have:

$$\boxed{\text{1-D delta method: } \mathrm{Var}_\theta(f(\hat{\theta}_n)) \approx f'(\theta)^2\, \mathrm{Var}_\theta(\hat{\theta}_n) \text{ if } \mathrm{Var}_\theta(\hat{\theta}_n) \text{ is small}}$$

Similarly, for $f: R^{d_1} \to R^{d_2}$ and $\nabla f(\theta) \in R^{d_1 \times d_2}$ defined as

$$(\nabla f(\theta))_{ij} = \frac{\partial}{\partial \theta_i} f_j, \quad 1 \le i \le d_1, \; 1 \le j \le d_2.$$

then

$$\boxed{\text{General delta method: } \; \mathrm{Cov}_\theta(f(\hat\theta_n)) \approx \nabla f(\theta)^T \mathrm{Cov}_\theta(\hat\theta_n) \nabla f(\theta) \quad \text{if } \|\mathrm{Cov}_\theta(\hat\theta_n)\| \text{ is small}}$$

3. Approximate variance for $\hat\theta_n$: by delta method.

$$\frac{1}{n} \nabla^2 A(\theta) = \mathrm{Cov}_\theta(\nabla A(\hat\theta_n)) \approx \nabla^2 A(\theta) \; \mathrm{Cov}_\theta(\hat\theta_n) \nabla^2 A(\theta)$$

$$\Longrightarrow \quad \boxed{\mathrm{Cov}_\theta(\hat\theta_n) \approx \frac{1}{n} (\nabla^2 A(\theta))^{-1} \approx \frac{1}{n} (\nabla^2 A(\hat\theta_n))^{-1}}$$

4. Practical way for variance estimation: bootstrap

Central idea of bootstrap: in order to estimate $\theta(P)$, one may use $\theta(P) \approx \theta(\hat P)$, with $\hat P$ typically being the empirical distribution.

In our case, $\theta(P)$ = variance of MLE based on $y_1, \cdots, y_n \sim P$

- if we knew P, we could resample m times from P (say $m = 1{,}000$):
  1) draw $y_1^{(i)}, y_2^{(i)}, \cdots, y_n^{(i)} \sim P$,
  2) compute the MLE $\hat\theta_n^{(i)}$ from $(y_1^{(i)}, \cdots, y_n^{(i)})$;
  3) compute the sample variance of $(\hat\theta_n^{(1)}, \cdots, \hat\theta_n^{(m)})$.

- however, we don't know P. Instead, we know $\hat P = \mathrm{unif}(\{y_1, \cdots, y_n\})$, the empirical distribution of n samples.

- computation of $\theta(\hat{P})$:
  1) draw $y_1^{(i)}, y_2^{(i)}, \cdots, y_n^{(i)} \sim \hat{P}$ (i.e. sample from $\{y_1, \cdots, y_n\}$ with replacement);
  2) compute the MLE $\hat{\theta}_n^{(i)}$ from $(y_1^{(i)}, \cdots, y_n^{(i)})$;
  3) compute the sample variance of $(\hat{\theta}_n^{(1)}, \cdots, \hat{\theta}_n^{(m)})$.

Some comments on bootstrap:

- bootstrap can be thought of as a general "plug-in" method;
- for example, if $Cov_\theta(\nabla A(\hat{\theta}_n)) = \frac{1}{n}\nabla^2 A(\theta)$ for some tractable $\nabla^2 A(\cdot)$, then a simple plug-in method is to use $\frac{1}{n}\nabla^2 A(\theta) \approx \frac{1}{n}\nabla^2 A(\hat{\theta}_n)$;
- however, if the computation of $\nabla^2 A(\cdot)$ is intractable, we can do:
  a) nonparametric bootstrap: sample $y_1^{(i)}, \cdots, y_n^{(i)} \sim \text{unif}\{y_1, \cdots, y_n\}$;
  b) parametric bootstrap: sample $y_1^{(i)}, \cdots, y_n^{(i)} \sim p_{\hat{\theta}_n}(y)$.

Example: Fisher's 2×2 table

R. A. Fisher considered the conditional distribution of X_1 given the row & column sums, i.e. (N, r_1, c_1):

	success	failure	
treatment	X_1 π_1	X_2 π_2	r_1
control	X_3 π_3	X_4 π_4	r_2
	c_1	c_2	N

$$p(x_1 \mid N, r_1, c_1) \propto \frac{N!}{x_1!\,(r_1-x_1)!\,(c_1-x_1)!\,(N-r_1-c_1+x_1)!}\,\pi_1^{x_1}\pi_2^{r_1-x_1}\pi_3^{c_1-x_1}\pi_4^{N-r_1-c_1+x_1}$$

$$\propto \frac{1}{x_1!\,(r_1-x_1)!\,(c_1-x_1)!\,(N-r_1-c_1+x_1)!}\underbrace{\left(\frac{\pi_1\pi_4}{\pi_2\pi_3}\right)^{x_1}}_{e^{\theta x_1}}$$

log odds: $\theta = \log\left(\frac{\pi_1\pi_4}{\pi_2\pi_3}\right)$ ($\theta = 0$: no treatment effect)

log-partition function. $A(\theta) = \log \sum_{x_1} \frac{e^{\theta x_1}}{x_1!\,(r_1-x_1)!\,(c_1-x_1)!\,(N-r_1-c_1+x_1)!}$.

The ulcdata is on the right.

	success	failure	
treatment	9 π_1	12 π_2	21
control	7 π_3	17 π_4	24
	16	29	45

Numerically one may evaluate:

- $\hat{\theta} = 0.600$
- $A''(\hat{\theta}) = 2.56$

$$\Rightarrow \mathrm{Var}(\hat{\theta}) \approx \frac{1}{A''(\hat{\theta})} = 0.391.$$

Question: how would you estimate $\mathrm{Var}(\hat{\theta})$ via bootstrap?

Inference of θ. $H_0.\ \theta = \theta_0$ vs. $H_1\ \theta \neq \theta_0$

1. 1-D inference ($\theta \in \mathbb{R}$)

- Pearson residual: $\frac{1}{n}\sum_{i=1}^{n} T(y_i) \rightsquigarrow N\left(A'(\theta), \frac{A''(\theta)}{n}\right)$

$$\boxed{R_p = \frac{\frac{1}{n}\sum_{i=1}^{n} T(y_i) - A'(\theta_0)}{\sqrt{A''(\theta_0)/n}} \xrightarrow[]{n\to\infty} N(0,1)}$$

- Deviance:

$$\boxed{\begin{aligned} D(\theta_1 ; \theta_2) &= 2\,\mathbb{E}_{\theta_1}\left[\log \frac{p_{\theta_1}(y)}{p_{\theta_2}(y)}\right] \\ &= 2\left(A(\theta_2) - A(\theta_1) - (\theta_2 - \theta_1)A'(\theta_1)\right) \geq 0 \end{aligned}}$$

Pf of second identity:

$$\mathbb{E}_{\theta_1}\left[\log \frac{p_{\theta_1}(y)}{p_{\theta_2}(y)}\right] = \mathbb{E}_{\theta_1}\left[(\theta_1 - \theta_2)T(y) - A(\theta_1) + A(\theta_2)\right]$$

$$= A(\theta_2) - A(\theta_1) - (\theta_2 - \theta_1)A'(\theta_1)$$

- deviance residual:

$$\boxed{R_D = \sqrt{n D(\hat{\theta}_n; \theta_0)}\ \mathrm{sign}\left(\frac{1}{n}\sum_{i=1}^{n} T(y_i) - A'(\theta_0)\right) \xrightarrow{n\to\infty} N(0,1)}$$

Intuition: $D(\hat{\theta}_n; \theta_0) = 2(A(\theta_0) - A(\hat{\theta}_n) - (\theta_0 - \hat{\theta}_n)A'(\hat{\theta}_n))$

$$\approx A''(\theta_0)\underbrace{(\hat{\theta}_n - \theta_0)^2}_{\approx \frac{1}{nA''(\theta_0)} Z^2 \text{ with } Z\sim N(0,1)}$$

- comparison of Pearson / deviance residuals: see HW2.

2. Multivariate inference ($\theta \in \mathbb{R}^d$)

- Wald test: $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{n\to\infty} N(0, \nabla^2 A(\theta_0)^{-1})$ under H_0.

$$\boxed{T_{n,\mathrm{Wald}} = n(\hat{\theta}_n - \theta_0)^T \nabla^2 A(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{n\to\infty} \chi^2_d}$$

- Rao's test (score test):

$$\sqrt{n}(\nabla A(\hat{\theta}_n) - \nabla A(\theta_0)) \xrightarrow{n\to\infty} N(0, \nabla^2 A(\theta_0)) \text{ under } H_0$$

$$\boxed{\begin{aligned} T_{n,\mathrm{Score}} &= n(\nabla A(\hat{\theta}_n) - \nabla A(\theta_0))^T \nabla^2 A(\theta_0)^{-1}(\nabla A(\hat{\theta}_n) - \nabla A(\theta_0)) \\ &= n\left(\frac{1}{n}\sum_{i=1}^{n} T(y_i) - \nabla A(\theta_0)\right)^T \nabla^2 A(\theta_0)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} T(y_i) - \nabla A(\theta_0)\right) \\ &\xrightarrow{n\to\infty} \chi^2_d \end{aligned}}$$

- Hoeffding's formula; deviance

$$D(\theta_1; \theta_2) = 2(A(\theta_2) - A(\theta_1) - \langle \theta_2 - \theta_1, \nabla A(\theta_1)\rangle)$$

$$\boxed{\begin{aligned} &\text{If } \hat{\theta}_n \text{ is the MLE based on } (y_1,\cdots,y_n), \text{ then for every } \theta, \\ &nD(\hat{\theta}_n; \theta) = 2\log\frac{p_{\hat{\theta}_n}(y_1,\cdots,y_n)}{p_\theta(y_1,\cdots,y_n)} \quad \text{(Pf: HW2)} \end{aligned}}$$

- likelihood ratio test:

$$\boxed{T_{n,LRT} = 2\log\frac{p_{\hat{\theta}_n}(y_1,\cdots,y_n)}{p_{\theta_0}(y_1,\cdots,y_n)} = nD(\hat{\theta}_n;\theta_0) \xrightarrow[n\to\infty]{} \chi^2_d \text{ under } H_0}$$

(known as Wilks' Theorem)

Intuition: $nD(\hat{\theta}_n;\theta_0) = 2n\left(A(\theta_0) - A(\hat{\theta}_n) - \langle \theta_0 - \hat{\theta}_n, \nabla A(\hat{\theta}_n)\rangle\right)$

$$\approx n(\theta_0 - \hat{\theta}_n)^T \nabla^2 A(\theta_0)(\theta_0 - \hat{\theta}_n)^T$$

$$= T_{n,Wald} \xrightarrow[n\to\infty]{} \chi^2_d.$$

3. Generalization to $H_0: \theta \in \Theta_0$ with $\dim(\Theta_0) = s < d$

Replace θ_0 by $\hat{\theta}_{0,n} = \underset{\theta\in\Theta_0}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i)$, then

$$T_{n,Wald},\ T_{n,Score},\ T_{n,LRT} \xrightarrow[n\to\infty]{} \chi^2_{d-s}.$$