STAT201B Lecture 7 Parametric Inference

1 Parametric Inference 概述

1.1 Definition: Location scale family

Logic

关于 location scale family 的更多描述, 见 STA3020 Lecture 3

令:

- Y 为服从分布 F 的随机变量
- F_μ 为 $Y+\mu$ 的 distribution function
- F_σ 为 σY 的 distribution function
- $F_{\mu,\sigma}$ 为 $\sigma Y+\mu$ 的 distribution function

则

- Family $\{F_\mu : -\infty < \mu < \infty\}$ 被称为 location family (e.g. $\mathcal{N}(\mu, 1)$)
- Family $\{F_\sigma : \sigma > 0\}$ 被称为 scale family (e.g. $\mathcal{N}(0, \sigma^2)$)
- Family $\{F_{\mu,\sigma} : -\infty < \mu < \infty, \sigma > 0\}$ 被称为 location scale family (e.g. $\mathcal{N}(\mu, \sigma^2)$)

Remark

WLOG, 我们通常假设 $\mathbb{E}[Y] = 0$, $Var[Y] = 1$

1.2 Definition: Parametric model

一个 parametric model 通常有以下形式:

$$\mathcal{F} = \{F(x, \theta) : \theta \in \Theta\}$$

其中 $\Theta \subset \mathbb{R}^k$ 为 parametric space

Remark

Class $\mathcal{F}$ 的选取通常基于我们对于特定问题的 knowledge (如 data generating mechanism), 需要特别注意是否存在违背这些 assumptions 的情形

Logic

接下来我们将介绍两种 parametric estimation methods: Method of Moment 和 Maximum Likelihood Estimation

2 Method of Moments 的定义

2.1 Definition: Method of Moments

Logic

关于 method of moments 的更多论述, 见 STA2004 Lecture 4

令:

- parameter of interest 为 $\theta = (\theta_1, \dots, \theta_k)$
- j^{th} (population) moment 为

$$\alpha_j := \alpha_j(\theta) = \mathbb{E}_\theta[X^j] = \int x^j dF_\theta(x), \quad \text{for } j = 1, \dots, k$$

- j^{th} sample moment 为

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$$

则 **method of moments (MOM) estimator $\hat{\theta}_n$** 满足:

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ &\cdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k \end{aligned}$$

Remark: MOM generalization

除了考虑 $\alpha_j(\theta) = \mathbb{E}_\theta[X^j]$, 我们还可以转而去考虑 $\alpha_j(\theta) = \mathbb{E}_\theta[g(X)^j]$, 并且令 $\hat{\theta}_n$ 满足:

$$\alpha_j(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n g(X_i)^j, \quad \text{for } j = 1, \dots, k$$

3 Maximum likelihood estimator 的定义

Logic

关于 maximum likelihood estimator 的更多描述, 见 STA3020 Lecture 7, 包括:

- MLE 的定义
- MLE 的 consistency 及其 conditions
- MLE 的 CAN property 及其 conditions
- MLE 的 invariance property

关于 likelihood function 的变种, 见 STA3020 Lecture 8, STA3020 Lecture 9, STA3020 Lecture 10, STA3020 Lecture 11, 包括:

- Composite likelihood
- Quasi likelihood
- Profile likelihood
- Generalized profile likelihood

3.1 Definition: Maximum likelihood estimator

令 sample $X = \{X_1, \dots, X_n\}$ 有 distribution function $f(x|\theta), \theta \in \Theta$,
则

- **likelihood function** 被定义为:

$$\mathcal{L}_n(\theta) = f(X_1, \dots, X_n; \theta) = \prod_{i=1}^n f(X_i; \theta) \quad \text{for indenpendet data}$$

Remark

换言之, likelihood 为 data 的 joint density, 但是被视作一个关于 θ 的函数

- **log-likelihood function** 被定义为:

$$l_n(\theta) = log\mathcal{L}_n(\theta) = \sum_{i=1}^{n} log f(X_i;\theta)$$

- **maximum likelihood estimator (MLE)** 被定义为

$$\hat{\theta}_{MLE} = \underset{\theta\in\Theta}{arg\,max}\ \mathcal{L}_n(\theta|x) = \underset{\theta\in\Theta}{arg\,max}\ l_n(\theta|x)$$

> ⚠ **Remark**
>
> - 若 log-likelihood 关于 θ differentiable, 则 MLE 的 candidates (in the interior of Θ) 满足:
>
> $$\frac{\partial}{\partial\theta_j} l_n(\theta) = 0, j = 1, \ldots, k$$
>
> 需要特别注意:
>   - **是否为 global maximum** (检查 second derivative)
>   - **maximum 是否位于 Θ 的 boundary** (first derivative 可能不为 0)
> - 可能会出现无法求出解析解的情况, 此时需要使用 numerical maximization methods
> - 该定义下的 MLE 不一定存在, 例如: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Unif(0,\theta)$, 则
>
> $$\mathcal{L}_n(\theta|x) = \frac{1}{\theta^n} \cdot \mathbf{1}\{max_{1\le i\le n}\{x_i\} < \theta\}$$
>
> 注意到 $\theta \in \Theta = \left(x_{(n)}, \infty\right)$, 因此实际上 θ 取不到 $x_{(n)}$; 为了避免这种情况, 可以将 MLE 定义在 Θ 的 closure 上:
>
> $$\hat{\theta}_{MLE} = \underset{\theta\in\bar{\Theta}}{arg\,max}\ \mathcal{L}_n(\theta|x) = \underset{\theta\in\bar{\Theta}}{arg\,max}\ l_n(\theta|x)$$

Example

问题:

令 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$, 求 θ 的 MLE

求解:

log-likelihood 为:

$$l(\theta) = \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi}}\right) + \left(-\frac{\sum_{i=1}^{n}(X_i-\theta)^2}{2}\right), \theta \in \mathbb{R}$$

求导, 可以得到:

$$\frac{\partial\, l(\theta)}{\partial\theta} = -\frac{\sum_{i=1}^{n} -2(X_i-\theta)}{2} = \sum_{i=1}^{n} X_i - n\theta$$

$$\frac{\partial^2 l(\theta)}{\partial\theta^2} = -n < 0$$

因此 $\hat{\theta}_{n,MLE} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$ 为 θ 的 MLE

Example

问题:

考虑同样的例子, 但是限制 $\Theta = [0, \infty)$

求解:

类似的, 可以得到:

$$
\begin{aligned}
l(\theta) &= \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi}}\right) + \left(-\frac{\sum_{i=1}^{n}(X_i-\theta)^2}{2}\right) \\
&:= C - \frac{\sum_{i=1}^{n} X_i^2 - 2\left(\sum_{i=1}^{n} X_i\right)\theta + n\theta^2}{2} \\
&= C - \frac{n(\theta^2 - 2\bar{X}\theta) + \sum_{i=1}^{n} X_i^2}{2} \\
&= C - \frac{n(\theta^2 - \bar{X}_n)^2 + \sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2}{2} \\
&:= C' - \frac{n(\theta - \bar{X}_n)}{2}
\end{aligned}
$$

因此可以得到:

$$
\hat{\theta}_n = \begin{cases} \bar{X} & \text{if } \bar{X} \geq 0 \\ 0 & \text{if } \bar{X} < 0 \end{cases} = max\{\bar{X}_n, 0\}
$$

Example

问题:
令 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Unif(0, \theta)$, 求 θ 的 MLE

求解:
Likelihood 为:

$$
\mathcal{L}(\theta) = \prod_{i=1}^{n}\left[\frac{1}{\theta} \cdot \mathbf{1}(X_i \geq 0)\mathbf{1}(X_i \leq \theta)\right] = \left(\frac{1}{\theta}\right)^n \cdot \mathbf{1}(X_{(1)} \geq 0) \cdot \mathbf{1}(X_{(n)} \leq \theta)
$$

若 $\hat{\theta}_n$ maximizes $\mathcal{L}_n(\theta)$, 则必须满足 $\mathbf{1}(X_{(n)} \leq \hat{\theta}_n) = 1$, 因此 $\hat{\theta}_n \geq X_{(n)}$;
与此同时, 为了 maximize $\left(\frac{1}{\theta}\right)^n$, 我们希望 $\hat{\theta}_n$ 尽可能的小, 因此 $\hat{\theta}_{n,MLE} = X_{(n)}$

Remark: θ 的 MoM

X 的 first moment 为: $\alpha_1(\theta) = \mathbb{E}[X_1] = \frac{\theta}{2}$,
X 的 first moment 的 plug-in estimator 为: $\hat{\alpha}_1 = \sum_{i=1}^{n} X_i/n$;
为了得到 $\hat{\theta}_{MoM}$, 需要求解: $\alpha_1(\hat{\theta}_{MoM}) = \hat{\alpha}_1$, 即

$$
\frac{\hat{\theta}_{MoM}}{2} = \sum_{i=1}^{n} X_i/n
$$

因此,

$$
\hat{\theta}_{MoM} = 2\bar{X}_n
$$

Example>

问题:
令 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Unif(\theta, \theta+1)$, 求 θ 的 MLE

求解:
Likelihood 为:

$$
\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i=1}^{n}[1 \cdot \mathbf{1}(X_i \geq \theta) \cdot \mathbf{1}(X_i \leq \theta + 1)] \\
&= \mathbf{1}(X_i \geq \theta) \cdot \mathbf{1}(X_i \leq \theta + 1) \\
&= \begin{cases} 1 & \text{if } X_{(n)} - 1 \leq \hat{\theta}_n \leq X_{(1)} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

因此 $\hat{\theta}_n$ 可以为 any value in $[X_{(n)} - 1, X_{(1)}]$

⚠ **Remark**

由此可以得出: MLE 不一定唯一

⚠ **Remark: MLE 是关于 sufficient statistics 的函数**

根据 Neyman factorization theorem:

$$T \text{ is sufficient for } \theta \iff \exists \text{ functions } h \text{ and } g, \text{ such that } p(x;\theta) = h(x) \cdot g(T(x), \theta)$$

MLE 可以写作:

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{arg\,max}\ \mathcal{L}_n(\theta|x) = \underset{\theta \in \Theta}{arg\,max}\ h(x) \cdot g(T(x), \theta) = \underset{\theta \in \Theta}{arg\,max}\ g(T(x), \theta)$$

仅与 sufficient statistics $T(x)$ 有关

4 Maximum Likelihood Estimator 的性质

Logic

接下来我们将分别介绍 MLE 的几个重要性质 (更多论述见 STA3020 Lecture 7):

1. Equivariance (invariance)
2. Consistency
3. Asymptotic normality
4. Asymptotic efficiency

对于后三个性质, 我们主要考虑 $\theta \in \Theta \subset \mathbb{R}$ 的情况

4.1 Definition: MLE 的 consistency condition

1. **independent distributed:** $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x;\theta)$
2. **identifiability:** 若 $\theta \neq \theta'$, 则 $f(x;\theta) \neq f(x;\theta')$
3. **common support:** Support $\{x : f(x;\theta) > 0\}$ 与 θ 的取值无关
4. **interior:** parameter space Θ 包含一个 open set ω, true parameter value θ^* 为其 interior point
5. **differentiable:** function $f(x;\theta)$ 在 ω 内部关于 θ differentiable

⚠ **Remark**

在 STA3020 Lecture 7 中, consistency condition 为:

1. **Separation condition:** Θ 为 compact 或 Θ 满足 $\underset{\theta \in \Theta, |\theta - \theta^*| \geq \epsilon}{\sup} \mathbb{E}[l(\theta)] < \mathbb{E}[l(\theta^*)], \quad \forall \epsilon > 0$ ($\mathbb{E}[l(\theta_0)]$ 严格大于其他 $\mathbb{E}[l(\theta)]$ 的 supremum)
2. **Convergence condition:** $\frac{1}{n} l_n(\theta)$ convergences uniformly to $\frac{1}{n}\mathbb{E}[l_n(\theta)]$ in probability, 即

$$\sup_{\theta \in \Theta} \frac{1}{n} |l_n(\theta) - \mathbb{E}[l_n(\theta)]| \xrightarrow{p} 0$$

STAT201B 中给出的 consistency condition 可以推出 convergence condition

These conditions ensure uniform convergence in probability of a normalized form of the log-likelihood to its expected value.

Note that

$$\begin{aligned}\ell_n(\theta) &= \sum_{i=1}^n \log f(X_i;\theta) \\ &\propto \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta) \\ &\xrightarrow{P} E_{\theta^*}[\log f(X_1;\theta)] \text{ for any fixed } \theta \text{ by WLLN}\end{aligned}$$

where θ^* denotes the true value of θ. Showing consistency requires that the convergence is uniform in θ. We also need to show that $E_{\theta^*}[\log f(X_1;\theta)]$ is maximized at $\theta = \theta^*$.

4.2 Theorem: MLE 的 consistency

若 MLE 的 consistency condition 满足,
则 $\hat{\theta}_{MLE}$ 为 true value θ^* 的一个 consistent estimator, 即 $\hat{\theta}_{MLE} \xrightarrow{p} \theta^*$

Proof

Proof of Theorem.1.3. Notice that,

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \ell(\theta) = \arg\max_{\theta\in\Theta}\left[\frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i|\theta)}{f(X_i|\theta_0)}\right] \triangleq \arg\max_{\theta\in\Theta} M(\theta)$$

Since we have

$$M(\theta) = \frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i|\theta)}{f(X_i|\theta_0)} \to \mathbb{E}\left[\log\frac{f(X_i|\theta)}{f(X_i|\theta_0)}\right] = -KL(f_{\theta_0}\|f_\theta) < 0,$$

where the last inequality sign changes to the equal sign iff $f(x|\theta) \equiv f(x|\theta_0), a.s.$, i.e., $\theta = \theta_0$. Therefore, if Θ is compact, then for $\forall \epsilon > 0$, we have

$$\sup_{\theta\in\Theta, |\theta-\theta_0|\geq\epsilon} \mathbb{E}M(\theta) < \mathbb{E}M(\theta_0) = -KL(f_{\theta_0}\|f_{\theta_0}) = 0. \tag{4.1}$$

Similarly we can also conclude (4.1) under the separation condition (1.1). Now, if we denote

$$\delta = \mathbb{E}M(\theta_0) - \sup_{\theta\in\Theta,|\theta-\theta_0|\geq\epsilon} \mathbb{E}M(\theta) > 0,$$

Since $M(\theta)$ converges uniformly to $\mathbb{E}M(\theta)$ in probability according to (ii) of Condition.1.2. Therefore, there exists $N \in \mathbb{N}^+$ s.t. when $n \geq N$,

$$\begin{aligned}&\mathbb{P}\left(|\hat{\theta}_{MLE} - \theta_0| \geq \epsilon\right) = \mathbb{P}\left(\sup_{\theta\in\Theta,|\theta-\theta_0|\geq\epsilon} M(\theta) > M(\theta_0)\right) \\ \leq&\mathbb{P}\left(\sup_{\theta\in\Theta,|\theta-\theta_0|\geq\epsilon}\left[M(\theta) - \mathbb{E}M(\theta)\right] > \left[M(\theta_0) - \mathbb{E}M(\theta_0)\right] + \delta\right) \\ \leq&\mathbb{P}\left(\sup_{\theta\in\Theta,|\theta-\theta_0|\geq\epsilon}\left[M(\theta) - \mathbb{E}M(\theta)\right] > \frac{\delta}{2}\right) + \mathbb{P}\left(\left[M(\theta_0) - \mathbb{E}M(\theta_0)\right] > -\frac{\delta}{2}\right) \\ \leq&2\mathbb{P}\left(\sup_{\theta\in\Theta}|M(\theta) - \mathbb{E}M(\theta)| > \frac{\delta}{2}\right) \to 0.\end{aligned}$$

Hence $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$. □

4.3 Definition: $g(\theta)$ 的 log-likelihood

若 θ 的 log-likelihood function 为 $l_n(\theta)$
则对于任意 function $g(\theta)$, $g(\theta)$ 的 log-likelihood function 被定义为:

$$l_g(\phi) := \sup_{\theta\in\Theta,\, g(\theta)=\phi} l(\theta)$$

Remark

1. 换言之, $g(\theta)=\phi$ 时的 log-likelihood = 先限定 θ 满足 $g(\theta)=\phi$, 再 maximize $l(\theta)$
2. 这么定义是为了满足 MLE 的 equivariance (invariance) property

4.4 Theorem: MLE 的 equivariance

令:

1. $\tau=g(\theta)$ 为关于 θ 的函数
2. $\hat{\theta}_n$ 为 θ 的 MLE

则 $\hat{\tau}_n=g(\hat{\theta}_n)$ 为 τ 的 MLE

Proof: g 为 one-to-one mapping 时的证明

若 g 为 one-to-one, 则存在 inverse g^{-1}, 因此关于 τ 的 (induced) likelihood 可以被定义为 $\mathcal{L}^*(\tau)=\mathcal{L}(g^{-1}(\tau))$.

注意到对于任意 τ, 有

$$\mathcal{L}^*(\tau)=\mathcal{L}(g^{-1}(\tau))\leq\mathcal{L}(\hat{\theta}_n)=\mathcal{L}^*(g(\hat{\theta}_n))$$

因此 $\hat{\tau}=g(\hat{\theta})$ maximizes $\mathcal{L}^*$

Proof: 更 general 的证明

Proof of Theorem.1.7. We proof by contradiction. If there exist some other $\theta_0\neq\hat{\theta}_{MLE}$ such that $\phi_0=g(\theta_0)\neq g(\hat{\theta}_{MLE})$, and $\phi_0=g(\theta_0)$ is the MLE of $g(\theta)$. Then

$$\ell_g(\phi_0)=\max_{\theta\in\Theta,g(\theta)=\phi_0}\ell(\theta)>\ell_g(g(\hat{\theta}_{MLE}))=\max_{\theta\in\Theta,g(\theta)=g(\hat{\theta}_{MLE})}\ell(\theta)=\ell(\hat{\theta}_{MLE}),$$

which contradict with the fact that $\hat{\theta}_{MLE}$ being the MLE of $\ell(\theta)$. Hence $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$. $\square$