Module 5: Statistical inference (II)

Siyue Yang

05/21/2022

Outline

This module we will review basics of parametric inference.

Parametric inference

Parametric models,

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$$

where the $\Theta \subset \mathbb{R}^k$ is the parameter space and $\theta = (\theta_1, \ldots, \theta_k)$ is the parameter.

- Goal of parametric inference: estimate the parametric θ (assume we known the form of the density).

Parameter of interest and nuisance parameter

Often, we are interested in estimating some function $T(\theta)$.

For example, if $X \sim N(\mu, \sigma^2)$, then

- Parameters: $\theta = (\mu, \sigma)$
- Parameter space: $\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$

If the goal is to estimate the μ then

- Parameter of interest: $T(\theta) = \mu$
- Nuisance parameter: σ

Methods for generating parametric estimators

1. Method of moments
2. Maximum likelihood

Method of moments

Suppose that the parameter $\theta = (\theta_1, \ldots, \theta_k)$ has k components.

- For $1 \leq j \leq k$, define the j^{th} moment

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}_\theta\left(X^j\right) = \int x^j dF_\theta(x)$$

- The j^{th} sample moment

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

- The method of moments estimator $\widehat{\theta}_n$

$$\alpha_1\left(\widehat{\theta}_n\right) = \widehat{\alpha}_1$$

$$\vdots$$

$$\alpha_k\left(\widehat{\theta}_n\right) = \widehat{\alpha}_k$$

Maximum likelihood

- Parametric model: $f(x;\theta)$, $X_1, \ldots, X_n$ iid
- Likelihood function

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f\left(X_i;\theta\right)$$

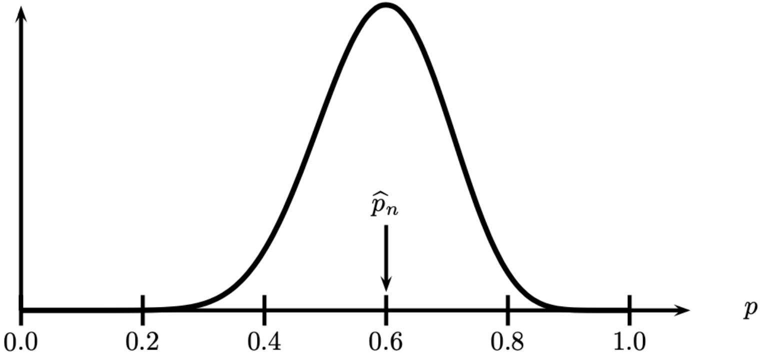
- The log-likelihood function

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

- The maximum likelihood estimator (MLE)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta)$$

An example of MLE



Likelihood function for Bernoulli with $n = 20$ and $\sum_{i=1}^{n} X_i = 12$.
The MLE is $\widehat{p}_n = 12/20 = 0.6$.

Why is maximum likelihood estimation so popular?

- A unified framework for estimation.
- Under mild regularity conditions, MLEs are
  1. **consistent** $\rightarrow$ converge to the true value in probability as $n \rightarrow \infty$, i.e.

     $$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| \leq \epsilon) = 1 \quad \forall \epsilon > 0$$

  2. **asymptotically normal** $\rightarrow \sqrt{n}(\hat{\theta} - \theta) \sim N\left(0, \sigma^2\right)$ for large n
  3. **asymptotically efficient** $\rightarrow$ achieve the lowest variance for large n
  4. **equivariant** $\rightarrow$ if $\hat{\theta}$ is the MLE for θ then $g(\hat{\theta})$ is the MLE for $g(\theta)$

Steps to find the MLE

1. Write out the likelihood

$$\mathcal{L}(\theta) = f\left(X_1, \ldots, X_n; \theta\right)$$

2. Simplify the log likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta)$$

3. Take the derivative of $\ell(\theta)$ with respect to the parameter of interest, θ Set $= 0$
4. Solve for θ (get $\hat{\theta}_{MLE}$)
5. Check that $\hat{\theta}_{MLE}$ is a maximum $\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta) < 0\right)$

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

1. The likelihood

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f\left(X_i; p\right) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

1. The likelihood

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$

2. Log-likelihood

$$\ell_n(p) = S \log p + (n - S) \log(1 - p)$$

Exercise

Suppose we have an iid sample $\{X_1, \ldots, X_n\}$ with $X_i \sim \text{Bernoulli}(p)$. Find the MLE for p.

1. The likelihood

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$

2. Log-likelihood

$$\ell_n(p) = S \log p + (n - S) \log(1 - p)$$

3. MLE (Solved the scoring equation)

$$\ell_n'(p) = 0$$

The MLE is $\widehat{p}_n = S/n$.

Score function and Fisher information

- Score function

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}$$

- Fisher information

$$\begin{aligned} I_n(\theta) &= \mathbb{V}_\theta \left(\sum_{i=1}^{n} s\left(X_i;\theta\right) \right) \\ &= \sum_{i=1}^{n} \mathbb{V}_\theta \left(s\left(X_i;\theta\right) \right) \end{aligned}$$

Asymptotic normality

Let $se = \sqrt{\mathbb{V}\left(\widehat{\theta}_n\right)}$. Under appropriate regularity conditions, the following hold:

1. se $\approx \sqrt{1/I_n(\theta)}$ and
$$\frac{\left(\widehat{\theta}_n - \theta\right)}{\text{se}} \rightsquigarrow N(0,1).$$
2. Let $\widehat{\text{se}} = \sqrt{1/I_n\left(\widehat{\theta}_n\right)}$. Then,
$$\frac{\left(\widehat{\theta}_n - \theta\right)}{\widehat{\text{se}}} \rightsquigarrow N(0,1)$$

- Let
$$C_n = \left(\widehat{\theta}_n - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta}_n + z_{\alpha/2}\widehat{\text{se}}\right)$$
Then, $\mathbb{P}_\theta\left(\theta \in C_n\right) \to 1-\alpha$ as $n \to \infty$.

Elements of likelihood estimation

One random variable: Given a model for X which assumes X has a density $f(x;\theta), \quad \theta \in \Theta \subset \mathbb{R}^k$, we have the following definitions:

likelihood function	$L(\theta;x) = c(x)f(x;\theta)$
log-likelihood function	$\ell(\theta;x) = \log L(\theta;x)$
score function	$u(\theta) = \partial\ell(\theta;x)/\partial\theta$
observed information function	$j(\theta) = -\partial^2\ell(\theta;x)/\partial\theta\partial\theta^T$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta\left\{U(\theta)U(\theta)^T\right\}$

Elements of likelihood estimation (i.i.d.)

Independent observations: When we have X_i independent, identically distributed from $f(x_i; \theta)$, then, denoting the observed sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ we have:

likelihood function	$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \theta)$
log-likelihood function	$\ell(\theta) = \ell(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \ell(\theta; x_i)$
maximum likelihood estimate	$\hat{\theta} = \hat{\theta}(\boldsymbol{x}) = \arg\sup_{\theta} \ell(\theta)$
score function	$U(\theta) = \ell'(\theta) = \sum U_i(\theta)$
observed information function	$j(\theta) = -\ell''(\theta) = -\ell''(\theta; \boldsymbol{x})$
observed (Fisher) information	$j(\hat{\theta})$
expected (Fisher) information	$i(\theta) = \mathrm{E}_{\theta}\left\{U(\theta)U(\theta)^T\right\} = n i_1(\theta)$

Delta method

If $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$ then

$$\frac{(\widehat{\tau}_n - \tau)}{\widehat{\mathsf{se}}(\widehat{\tau})} \rightsquigarrow N(0,1)$$

where $\widehat{\tau}_n = g\left(\widehat{\theta}_n\right)$ and

$$\widehat{\mathsf{se}}\left(\widehat{\tau}_n\right) = \left|g'(\widehat{\theta})\right| \widehat{\mathsf{se}}\left(\widehat{\theta}_n\right)$$

Hence, if

$$C_n = \left(\widehat{\tau}_n - z_{\alpha/2}\widehat{\mathsf{se}}\left(\widehat{\tau}_n\right), \widehat{\tau}_n + z_{\alpha/2}\widehat{\mathsf{se}}\left(\widehat{\tau}_n\right)\right)$$

then $\mathbb{P}_\theta\left(\tau \in C_n\right) \to 1 - \alpha$ as $n \to \infty$.

Exercise

Let $X_1, \ldots, X_n \sim$ Bernoulli(p) and let $\psi = g(p) = \log(p/(1- \ p)\ )$.

Exercise

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\psi = g(p) = \log(p/(1- \ p)\ )$.

The Fisher information function is $I(p) = 1/(p(1-p))$

Exercise

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\psi = g(p) = \log(p/(1- \ p)\ )$.

The Fisher information function is $I(p) = 1/(p(1-p))$

The estimated standard error of the MLE $\widehat{p}_n$ is

$$\widehat{\mathrm{se}} = \sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}$$

Exercise

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\psi = g(p) = \log(p/(1-\ p)\ )$.

The Fisher information function is $I(p) = 1/(p(1-p))$

The estimated standard error of the MLE $\widehat{p}_n$ is

$$\widehat{\mathrm{se}} = \sqrt{\frac{\widehat{p}_n\left(1-\widehat{p}_n\right)}{n}}$$

The MLE of ψ is $\widehat{\psi} = \log \widehat{p}/(1-\widehat{p})$. Since, $g'(p) = 1/(p(1-p))$, according to the delta method

$$\widehat{\mathrm{se}}\left(\widehat{\psi}_n\right) = \left|g'\left(\widehat{p}_n\right)\right| \widehat{\mathrm{se}}\left(\widehat{p}_n\right) = \frac{1}{\sqrt{n\widehat{p}_n\left(1-\widehat{p}_n\right)}}$$

Exercise

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\psi = g(p) = \log(p/(1- \ p) \ )$.

The Fisher information function is $I(p) = 1/(p(1-p))$

The estimated standard error of the MLE $\widehat{p}_n$ is

$$\widehat{\mathrm{se}} = \sqrt{\frac{\widehat{p}_n\left(1-\widehat{p}_n\right)}{n}}$$

The MLE of ψ is $\widehat{\psi} = \log \widehat{p}/(1-\widehat{p})$. Since, $g'(p) = 1/(p(1-p))$, according to the delta method

$$\widehat{\mathrm{se}}\left(\widehat{\psi}_n\right) = \left|g'\left(\widehat{p}_n\right)\right| \widehat{\mathrm{se}}\left(\widehat{p}_n\right) = \frac{1}{\sqrt{n\widehat{p}_n\left(1-\widehat{p}_n\right)}}$$

An approximate 95 percent confidence interval is

$$\widehat{\psi}_n \pm \frac{2}{\sqrt{n\widehat{p}_n\left(1-\widehat{p}_n\right)}}$$

MLE in R

Sometimes, there is no closed-form solution, so we need to use optimization methods to find the maximum of the log-likelihood.

- `optim()` find values of some parameters that **minimizes** some function.
- Newton-Raphson
- EM-algorithm

Newton-Raphson

Derivative of the log-likelihood around θ^3 :

$$0 = \ell'(\widehat{\theta}) \approx \ell'\left(\theta^j\right) + \left(\widehat{\theta} - \theta^j\right)\ell''\left(\theta^j\right)$$

Solving for $\widehat{\theta}$ gives

$$\widehat{\theta} \approx \theta^j - \frac{\ell'\left(\theta^j\right)}{\ell''\left(\theta^j\right)}.$$

This suggests the following iterative scheme:

$$\widehat{\theta}^{j+1} = \theta^j - \frac{\ell'\left(\theta^j\right)}{\ell''\left(\theta^j\right)}$$

In the multiparameter case, the mle $\widehat{\theta} = \left(\widehat{\theta}_1, \ldots, \widehat{\theta}_k\right)$ is a vector and the method becomes

$$\widehat{\theta}^{j+1} = \theta^j - H^{-1}\ell'\left(\theta^j\right)$$

where $\ell'\left(\theta^j\right)$ is the vector of first derivatives and H is the matrix of second derivatives of the log-likelihood.

Expectation-Maximization (EM) algorithm

Idea: Iterate between taking an expectation then maximizing.

Suppose we have data Y whose density $f(y;\theta)$ leads to a log-likelihood that is hard to maximize. However we can find another variable Z s.t. $f(y;\theta) = \int f(y,z;\theta)dz$ and $f(y,z;\theta)$ is easy to maximize.

- Pick a starting value θ^0. Now for $j = 1, 2, \ldots$, repeat steps E and M below:
- (The E-step): Calculate

$$J\left(\theta \mid \theta^j\right) = \mathbb{E}_{\theta^j}\left(\log \frac{f\left(Y^n, Z^n; \theta\right)}{f\left(Y^n, Z^n; \theta^j\right)} \mid Y^n = y^n\right).$$

  The expectation is over the missing data Z^n treating θ^i and the observed data Y^n as fixed.
- (M-step) Find θ^{j+1} to maximize $J\left(\theta \mid \theta^j\right)$