Method of moments and Maximum likelihood estimation

Botond Szabo

Leiden University

Leiden, 11. 03. 2019

Outline

1. Method of moments
2. Maximum likelihood
3. Asymptotic normality
4. Optimality
5. Delta method
6. Parametric bootstrap
7. Quiz

Method of moments

- Recall: the jth moment of the random variable $X \sim f_\theta$ is

$$\alpha_j(\theta) := E_\theta X^j = \int x^j f(x;\theta)dx.$$

- The jth sample moment is defined as

$$\hat{\alpha}_j := n^{-1} \sum_{i=1}^{n} X_i^j.$$

- For $\theta = (\theta_1, ...\theta_k)$ the method of moments estimator $\hat{\theta}_n$ is defined as

$$\alpha_1(\hat{\theta}) = \hat{\alpha}_1,$$

$$....$$

$$\alpha_k(\hat{\theta}) = \hat{\alpha}_k,$$

Examples I

Example

Let $X_1, \ldots, X_n \sim Bernoulli(p)$. Then $\alpha_1(p) = E_p X_1 = p$ and $\hat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$. By equating these we get

$$\hat{p}_n = n^{-1} \sum_{i=1}^n X_i.$$

Examples II

Example

Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$. Then $\alpha_1 = EX_1 = \mu$ and $\alpha_2 = EX_1^2 = \mu^2 + \sigma^2$. Furthermore $\hat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$ and $\hat{\alpha}_2 = n^{-1} \sum_{i=1}^n X_i^2$. Solving the equations

$$\hat{\mu}_n = n^{-1} \sum_{i=1}^n X_i, \quad \text{and} \quad \hat{\mu}^2 + \hat{\sigma}^2 = n^{-1} \sum_{i=1}^n X_i^2,$$

we get that

$$\hat{\mu}_n = n^{-1} \sum_{i=1}^n X_i, \quad \text{and} \quad \hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Properties

Theorem

Let $\hat{\theta}_n$ denote the method of moments estimator. Under appropriate conditions on the model, the following statements hold:

- *The estimate $\hat{\theta}_n$ exists with probability tending to one.*
- *The estimate is consistent , i.e. $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.*
- *The estimate is asymptotically normal (the precise form is given in the book).*

Likelihood

Definition

Let $X_1, \ldots, X_n$ be IID with PDF (or PMF) $f(x;\theta)$. The likelihood function is defined by

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

The log-likelihood function is defined by $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

Maximum likelihhod estimator

Definition

The maximum likelihood estimator (MLE), denoted by $\hat{\theta}_n$, is the value of θ that maximises $\mathcal{L}_n(\theta)$ (or, equivalently, $\log \mathcal{L}_n(\theta)$).

Example

Example

Suppose that $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The PMF is $f(x;p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. The unknown parameter is p. Then

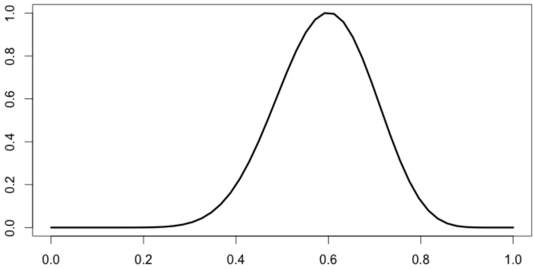
$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S},$$

where $S = \sum_{i=1}^{n} X_i$. Therefore,

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

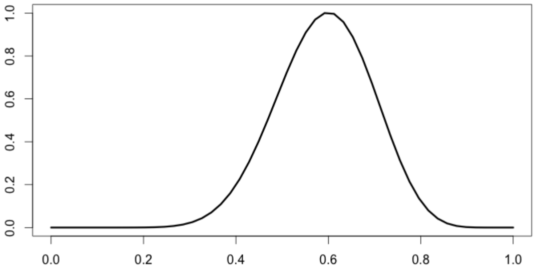
To find the maximum, take a derivative of $\ell_n(\theta)$, set it to zero and find that $\hat{p}_n = S/n$.

Example



Likelihood for Bernoulli with $n = 20$ and $\sum_{i=1}^{n} X_i = 12$.

Example



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

Example

Example

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. The unknown parameter is $\theta = (\mu, \sigma^2)$ and the likelihood (up to some constants not depending on θ) is

$$\begin{aligned}\mathcal{L}_n(\theta) &= \prod_{i=1}^n \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right) \\ &= \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\right) \\ &= \sigma^{-n} \exp\left(-\frac{nS^2}{2\sigma^2}\right) \exp\left(-\frac{n(\overline{X}_n - \mu)^2}{2\sigma^2}\right),\end{aligned}$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $S^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Example (continued)

Example

The log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X}_n - \mu)^2}{2\sigma^2}.$$

Solving the equations

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0, \quad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0,$$

we conclude that $\hat{\mu}_n = \overline{X}_n$ and $\hat{\sigma}_n = S$ (it can be verified that these are the global maxima of the likelihood).

Properties of maximum likelihood estimators

Theorem

Under suitable conditions on $f(x;\theta)$, the MLE $\hat{\theta}_n$ is consistent: $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.

Properties of maximum likelihood estimators

Theorem

Under suitable conditions on $f(x;\theta)$, the MLE $\hat{\theta}_n$ is consistent: $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.

Theorem

Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be an MLE of θ. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is an MLE of τ.

Properties of maximum likelihood estimators

Theorem

Under suitable conditions on $f(x;\theta)$, the MLE $\hat{\theta}_n$ is consistent: $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.

Theorem

Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be an MLE of θ. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is an MLE of τ.

Example

Let $X_1, \ldots, X_n \sim N(\theta, 1)$. The MLE for θ is $\hat{\theta}_n = \overline{X}_n$. Let $\tau = e^{\theta}$. Then the MLE for τ is $\hat{\tau}_n = e^{\overline{X}_n}$.

Score function and Fisher information

Definition

The score function is defined by

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

The Fisher information *in n IID observations $X_1, \ldots, X_n \sim f(x;\theta)$ is*

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

Properties

Theorem

For $n = 1$ write $I(\theta)$ instead of $I_1(\theta)$. One has $\mathbb{E}_\theta[s(X_1;\theta)] = 0$ and hence $\mathbb{V}_\theta(s(X_i;\theta)) = \mathbb{E}[s^2(X_1;\theta)]$.

Properties

Theorem

For $n = 1$ write $I(\theta)$ instead of $I_1(\theta)$. One has $\mathbb{E}_\theta[s(X_1;\theta)] = 0$ and hence $\mathbb{V}_\theta(s(X_i;\theta)) = \mathbb{E}[s^2(X_1;\theta)]$.

Theorem

One has $I_n(\theta) = nI(\theta)$. Also,

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right] \\ &= -\int \left(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right) f(x;\theta)dx. \end{aligned}$$

Asymptotic normality

Theorem

Let $\hat{\theta}_n$ *be an MLE and* $\mathrm{se} = \sqrt{\mathbb{V}_\theta[\hat{\theta}_n]}$. *Under appropriate conditions,*

1. $\mathrm{se} \approx \sqrt{1/I_n(\theta)}$ *and*

$$\frac{\hat{\theta}_n - \theta}{\mathrm{se}} \rightsquigarrow N(0,1).$$

2. *Let* $\hat{\mathrm{se}} = \sqrt{1/I_n(\hat{\theta}_n)}$. *Then*

$$\frac{\hat{\theta}_n - \theta}{\hat{\mathrm{se}}} \rightsquigarrow N(0,1).$$

Confidence interval

Theorem

Let

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{\text{se}}, \hat{\theta}_n + z_{\alpha/2}\hat{\text{se}}).$$

Then $\mathbb{P}_\theta(\theta \in C_n) \to 1 - \alpha$ *as* $n \to \infty$.

In particular, for $\alpha = 0.05$, $z_{\alpha/2} = 1.96 \approx 2$, *and*

$$\hat{\theta}_n \pm 2\hat{\text{se}}$$

is an approximate 95% *confidence interval.*

Example

Example

Let $X_1, \ldots, X_n$ be IID Poisson(λ). Then it can be shown that $\hat{\lambda}_n = \overline{X}_n$ and $I(\lambda) = 1/\lambda$, so that

$$\hat{\text{se}} = \frac{1}{\sqrt{nI(\hat{\lambda}_n)}} = \sqrt{\frac{\hat{\lambda}_n}{n}}.$$

Therefore, an approximate $1 - \alpha$ confidence interval for λ is

$$\hat{\lambda}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}_n}{n}}.$$

MLE vs. median

Example

Let $X_1, \ldots, X_n$ be IID $N(\theta, 1)$. The MLE for θ is $\hat{\theta}_n = \overline{X}_n$. Another reasonable estimator for θ is the sample median $\widetilde{\theta}_n$.

The MLE satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, 1).$$

The sample median can be shown to satisfy

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \rightsquigarrow N\left(0, \frac{\pi}{2}\right).$$

The sample median converges to the right value, but has a larger variance than MLE.

Efficiency of MLE

Theorem

Let $\hat{\theta}_n$ be an MLE and $\widetilde{\theta}_n$ (almost) any other estimator. Suppose

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \sigma^2_{\text{MLE}}), \quad \sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \sigma^2_{\text{tilde}}).$$

Define the asymptotic relative efficiency as

$$\text{ARE}(\widetilde{\theta}_n, \hat{\theta}_n) = \frac{\sigma^2_{\text{MLE}}}{\sigma^2_{\text{tilde}}}.$$

Then $\text{ARE}(\widetilde{\theta}_n, \hat{\theta}_n) \leq 1$. *Thus the MLE has the smallest (asymptotic) variance and we say that the MLE is optimal or asymptotically efficient.*

Delta method

Theorem

If $\tau = g(\theta)$, where g is differentiable with $g'(\theta) \neq 0$, then for $\hat{\tau}_n = g(\hat{\theta}_n)$,

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

Here

$$\hat{\mathsf{se}}(\hat{\tau}_n) = |g'(\hat{\theta}_n)|\hat{\mathsf{se}}(\hat{\theta}_n).$$

Hence, if

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

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

Example

Example

Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$. Suppose we want to estimate $\tau = \log \sigma$. The log-likelihood is

$$\ell(\sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} X_i^2.$$

Differentiate $\ell(\sigma)$ and set the derivative to zero to conclude that

$$\hat{\sigma}_n = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}}.$$

With some effort we can show that $I(\sigma) = 2/\sigma^2$. Therefore, $\widehat{\text{se}}(\hat{\sigma}_n) = \hat{\sigma}_n/\sqrt{2n}$.

Example (continued)

Example

Now $\hat{\tau}_n = \log \hat{\sigma}_n$. Since $(\log \sigma)' = 1/\sigma$, we get that

$$\widehat{\mathsf{se}}(\hat{\tau}_n) = \frac{1}{\hat{\sigma}_n} \frac{\hat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}},$$

and an approximate 95% confidence interval for τ is

$$\hat{\tau}_n \pm \sqrt{\frac{2}{n}}.$$

Parametric bootstarp

- The bootstrap is a method for estimating standard errors, computing confidence intervals and other quantities that might be difficult to compute otherwise.
- In the parametric models we know that $X_1, \dots, X_n \sim f(x; \theta)$.
- Suppose we want to approximate the standard error of $T(X_1, \dots, X_n)$, denoted by $\mathrm{se} = \sqrt{\mathbb{V}_\theta[T]}$.
- The idea of the parametric bootstrap is to approximate se with $\sqrt{\mathbb{V}_{\hat{\theta}_n}[T]}$, where $\hat{\theta}_n$ is e.g. the MLE. However, in many cases it is difficult to compute $\mathbb{V}_{\hat{\theta}_n}[T]$. Idea: approximate it via simulation.

Parametric bootstarp: sampling

- Sample $X_1^*, \ldots, X_n^*$ from $f(x; \hat{\theta}_n)$.
- Compute the estimator T using the "new sample" $X_1^*, \ldots, X_n^*$:

$$T^* = T(X_1^*, \ldots, X_n^*).$$

- Repeat this procedure B times resulting $T_1^*, \ldots, T_B^*$.
- Denote the average of the outomes by $\bar{T}^*$.
- Approximate $\mathbb{V}_{\hat{\theta}_n}[T]$ by

$$\hat{se}_{boot}^2 = B^{-1} \sum_{b=1}^{B} \left(T_b^* - \bar{T}^*\right)^2.$$

Example

Example

Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$. The MLE of σ is

$$\hat{\sigma}_n = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}.$$

We use the parametric bootstrap to approximate se.

Simulate $X_1^*, \ldots, X_n^* \sim N(0, \hat{\sigma}_n^2)$ and compute $\hat{\sigma}^* = \sqrt{n^{-1} \sum_{i=1}^n X_i^{*2}}$. Next repeat this process B times to get $\hat{\sigma}_1^*, \ldots, \hat{\sigma}_B^*$. Finally, set

$$\hat{\mathsf{se}}_{\mathrm{boot}} = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(\hat{\sigma}_b^* - \frac{1}{B} \sum_{j=1}^{B} \hat{\sigma}_j^* \right)^2}.$$

Question 1

In what sense is the MLE optimal?

Answers:

1. It has the smallest asymptotic variance.
2. It is unbiased.
3. It is a normally distributed random variable and therefore easy to work with.
4. It is the value minimizing the likelihood function.

Question 1

In what sense is the MLE optimal?

Answers:

1. It has the smallest asymptotic variance.
2. It is unbiased.
3. It is a normally distributed random variable and therefore easy to work with.
4. It is the value minimizing the likelihood function.

Question 2

What is the Delta method good for

Answers:

1. Compute the MLE for a function of the paramter $\tau = g(\theta)$.
2. Construct a confidence set for a function of the paramter $\tau = g(\theta)$.
3. Compute the Fisher information.
4. Compute the relative efficiency.

Question 2

What is the Delta method good for

Answers:

1. Compute the MLE for a function of the paramter $\tau = g(\theta)$.
2. Construct a confidence set for a function of the paramter $\tau = g(\theta)$.
3. Compute the Fisher information.
4. Compute the relative efficiency.

Question 3

If $\tau = g(\theta)$ and $\hat{\tau}_n = g(\hat{\theta}_n)$, then what is the estimated standard error of $\hat{\tau}_n$

Answers:

1. $\widehat{se}(\hat{\tau}_n) = \widehat{se}(\hat{\theta}_n)$.
2. $\widehat{se}(\hat{\tau}_n) = \widehat{se}(\hat{\theta}_n)/|g'(\hat{\theta}_n)|$.
3. $\widehat{se}(\hat{\tau}_n) = |g(\hat{\theta}_n)|\widehat{se}(\hat{\theta}_n)$.
4. $\widehat{se}(\hat{\tau}_n) = |g'(\hat{\theta}_n)|\widehat{se}(\hat{\theta}_n)$.

Question 3

If $\tau = g(\theta)$ and $\hat{\tau}_n = g(\hat{\theta}_n)$, then what is the estimated standard error of $\hat{\tau}_n$.

Answers:

1. $\hat{se}(\hat{\tau}_n) = \hat{se}(\hat{\theta}_n)$.
2. $\hat{se}(\hat{\tau}_n) = \hat{se}(\hat{\theta}_n)/|g'(\hat{\theta}_n)|$.
3. $\hat{se}(\hat{\tau}_n) = |g(\hat{\theta}_n)|\hat{se}(\hat{\theta}_n)$.
4. $\hat{se}(\hat{\tau}_n) = |g'(\hat{\theta}_n)|\hat{se}(\hat{\theta}_n)$.

Question 4

What is the bootstrap method good for?

Answers:

1. To estimate standard error and compute confidence intervals.
2. To help put on boots.
3. To compute the MLE.
4. Compute the Fisher information.

Question 4

What is the bootstrap method good for?

Answers:

1. To estimate standard error and compute confidence intervals.
2. To help put on boots.
3. To compute the MLE.
4. Compute the Fisher information.

Question 5

What does the star denote in $X_1^*, \ldots, X_n^*$ in the bootstrap method?

Answers:

1. They are a permutation of $X_1, \ldots, X_n$.
2. They are a "new" sample (with replacement) from the numbers $X_1, \ldots, X_n$.
3. They are a "new" sample (without replacement) from the numbers $X_1, \ldots, X_n$.
4. They are new draws from the distribution F.

Question 5

What does the star denote in $X_1^*, \ldots, X_n^*$ in the bootstrap method?

Answers:

1. They are a permutation of $X_1, \ldots, X_n$.
2. They are a "new" sample (with replacement) from the numbers $X_1, \ldots, X_n$.
3. They are a "new" sample (without replacement) from the numbers $X_1, \ldots, X_n$.
4. They are new draws from the distribution F.

Question 6

How do we estimate the variance using the bootstrap sample?

Answers:

1. $B^{-1} \sum_{b=1}^{B} \left(T_b^* - \bar{T}^*\right)^2.$
2. $B^{-1} \sum_{b=1}^{B} \left(X_b - \bar{T}\right)^2.$
3. $B^{-1} \sum_{b=1}^{B} \left(T_b^* - \bar{T}\right)^2.$
4. $B^{-1} \sum_{b=1}^{B} \left(X_b - \bar{T}^*\right)^2.$

Question 6

How do we estimate the variance using the bootstrap sample $\hat{se}_{boot}$?

Answers:

1. $B^{-1}\sum_{b=1}^{B}\left(T_b^* - \bar{T}^*\right)^2.$
2. $B^{-1}\sum_{b=1}^{B}\left(X_b - \bar{T}\right)^2.$
3. $B^{-1}\sum_{b=1}^{B}\left(T_b^* - \bar{T}\right)^2.$
4. $B^{-1}\sum_{b=1}^{B}\left(X_b - \bar{T}^*\right)^2.$