Statistics

MATH 324

McGill University, Montréal, Canada

Fall 2018

Introduction

In this section we will discuss two systematic ways of deriving point estimation(s) of parameters in a parametric family.

(1) Method of moments

(2) Method of maximum likelihood

Sections 9.6-9.8

A question:

- Let $X_1, X_2, \ldots, X_n$ be an iid sample from a parametric family

$$\mathcal{F} = \{F(\cdot;\theta); \theta \in \Theta\}$$

- This means, we know $F(\cdot;\theta)$ up to an unknown parameter θ:

Normal, Poisson, Binomial, ...

- Question:

Given the sample, how to estimate θ?

A question:

- Let $X_1, X_2, \ldots, X_n$ be an iid sample from a parametric family

$$\mathcal{F} = \{F(\cdot;\theta); \theta \in \Theta\}$$

- This means, we know $F(\cdot;\theta)$ up to an unknown parameter θ:

Normal, Poisson, Binomial, ...

- Question:

Given the sample, how to estimate θ?

A question:

- Let $X_1, X_2, \ldots, X_n$ be an iid sample from a parametric family

$$\mathcal{F} = \{F(\cdot;\theta); \theta \in \Theta\}$$

- This means, we know $F(\cdot;\theta)$ up to an unknown parameter θ:

Normal, Poisson, Binomial, ...

- Question:

Given the sample, how to estimate θ?

What we have discussed so far:

We saw examples of parameter estimators and concluded that:

- An estimator $\hat{\theta}_n$ should be unbiased; at least asymptotically.
- Its MSE should be small.
- It should be consistent.
- A minimum variance unbiased estimator (if exists) can (in principle) be constructed from a sufficient statistic.
- **We need a systematic and feasible way to derive "good" estimators.**

I. The method of moments:

- This method was introduced by Karl Pearson.

- In this method, we basically match the “sample” and “population” methods and obtain the parameter estimates.

Population and sample moments

- Consider a random variable X with a distribution $F(\cdot;\theta)$. For $k \in \mathbb{N}$, we have that (if it exists)

$$E(X^k) = \begin{cases} \sum_x x^k f(x;\theta) & , X \text{ discrete;} \\ \int_{-\infty}^{\infty} x^k \; f(x \; \theta) dx & , X \text{ continuous.} \end{cases}$$

are the k-th moments of X.

- Based on a random sample $X_1, \ldots, X_n$, the sample moments are

$$m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Method of moments: (Karl Pearson)

- Definition:

  If d parameters are unknown, we estimate them by solving the d equations

$$m_k = E(X^k) \ , \ \ k = 1, 2, \ldots, d$$

  The resulting estimators are called moment estimators.

Examples

- We will discuss examples in class.

Summary

Our observations from the examples:

(1) The moment estimators are:

- easy to compute for most of the parametric families.
- typically consistent.

(2) However, the moment estimators may

- be biased and hence not MVUE; Examples 4 and 6
- be inadmissible; Example 4
- behave badly; Example 7

The method of maximum likelihood

- The method was designed by Sir R.A. Fisher in the 1910s. It is the most popular and effective estimation method in statistics.

The likelihood function

- Definition 9.4:

  Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a parametric family $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$, where Θ is the parameter space which denotes the set of all admissible parameter values. Let $x_1, x_2, \ldots, x_n$ be the observed values of the sample. The likelihood function of θ is defined by

$$L_n(\theta) = f(x_1;\theta) \times f(x_2;\theta) \times \ldots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

- The log-likelihood function of θ is given by:

$$l_n(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

The likelihood function

- Definition 9.4:

  Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a parametric family $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$, where Θ is the parameter space which denotes the set of all admissible parameter values. Let $x_1, x_2, \ldots, x_n$ be the observed values of the sample. The likelihood function of θ is defined by

$$L_n(\theta) = f(x_1;\theta) \times f(x_2;\theta) \times \ldots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

- The log-likelihood function of θ is given by:

$$l_n(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

Remarks

- When X is discrete, the likelihood function is exactly the probability of observing what we have observed as $x_1, x_2, \ldots, x_n$.
- When X is continuous, the likelihood function is approximately proportional to the probability of observing what we have observed as $x_1, x_2, \ldots, x_n$.
- The likelihood function is regarded as a deterministic real-valued function of the parameter θ.
- Recall: we used the likelihood function in the Fisher-Neyman Factorization Theorem to obtain sufficient statistic(s) for the corresponding parametric family.

Remarks

- When X is discrete, the likelihood function is exactly the probability of observing what we have observed as $x_1, x_2, \ldots, x_n$.
- When X is continuous, the likelihood function is approximately proportional to the probability of observing what we have observed as $x_1, x_2, \ldots, x_n$.
- The likelihood function is regarded as a deterministic real-valued function of the parameter θ.
- Recall: we used the likelihood function in the Fisher-Neyman Factorization Theorem to obtain sufficient statistic(s) for the corresponding parametric family.

Remarks

- When X is discrete, the likelihood function is exactly the probability of observing what we have observed as $x_1, x_2, \ldots, x_n$.
- When X is continuous, the likelihood function is approximately proportional to the probability of observing what we have observed as $x_1, x_2, \ldots, x_n$.
- The likelihood function is regarded as a deterministic real-valued function of the parameter θ.
- Recall: we used the likelihood function in the Fisher-Neyman Factorization Theorem to obtain sufficient statistic(s) for the corresponding parametric family.

Motivation

- In the method of maximum likelihood, we estimate the parameter of interest by obtaining a value of θ that maximizes $L_n(\theta)$.

- That is, we obtain a value of θ that maximizes the probability of observing what we have observed as our data.

- Thus, it makes sense to estimate θ by

$$\hat{\theta}_n = argmax_{\theta \in \Theta} \; L_n(\theta).$$

and note that $\hat{\theta}_n \in \Theta$.

Motivation

- In the method of maximum likelihood, we estimate the parameter of interest by obtaining a value of θ that maximizes $L_n(\theta)$.
- That is, we obtain a value of θ that maximizes the probability of observing what we have observed as our data.
- Thus, it makes sense to estimate θ by

$$\hat{\theta}_n = argmax_{\theta \in \Theta} \, L_n(\theta).$$

and note that $\hat{\theta}_n \in \Theta$.

Motivation

- In the method of maximum likelihood, we estimate the parameter of interest by obtaining a value of θ that maximizes $L_n(\theta)$.

- That is, we obtain a value of θ that maximizes the probability of observing what we have observed as our data.

- Thus, it makes sense to estimate θ by

$$\hat{\theta}_n = argmax_{\theta \in \Theta} \, L_n(\theta).$$

and note that $\hat{\theta}_n \in \Theta$.

Motivation

- In the method of maximum likelihood, we estimate the parameter of interest by obtaining a value of θ that maximizes $L_n(\theta)$.
- That is, we obtain a value of θ that maximizes the probability of observing what we have observed as our data.
- Thus, it makes sense to estimate θ by

$$\hat{\theta}_n = argmax_{\theta \in \Theta} \; L_n(\theta).$$

and note that $\hat{\theta}_n \in \Theta$.

Maximum likelihood estimate (MLE)

- Defintion:

  Suppose $x_1, x_2, \ldots, x_n$ is the observed values of a random sample from a parametric family $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, where Θ is the parameter space which denotes the set of all admissible values of the parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$.

  The maximum likelihood estimate of θ is given by

$$\hat{\theta}_n = argmax_{\theta\in\Theta} \;\; L_n(\theta).$$

- We assume that this maximum is unique; it is often, but not always, the case in practice.

Maximum likelihood estimate (MLE)

- Defintion:

  Suppose $x_1, x_2, \ldots, x_n$ is the observed values of a random sample from a parametric family $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, where Θ is the parameter space which denotes the set of all admissible values of the parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$.

  The maximum likelihood estimate of θ is given by

$$\hat{\theta}_n = argmax_{\theta\in\Theta} \; L_n(\theta).$$

- We assume that this maximum is unique; it is often, but not always, the case in practice.

Maximum likelihood estimate (MLE)

- Defintion:

  Suppose $x_1, x_2, \ldots, x_n$ is the observed values of a random sample from a parametric family $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, where Θ is the parameter space which denotes the set of all admissible values of the parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$.

  The maximum likelihood estimate of θ is given by

  $$\hat{\theta}_n = argmax_{\theta\in\Theta} \;\; L_n(\theta).$$

- We assume that this maximum is unique; it is often, but not always, the case in practice.

Remark

- It is often much easier to work with the log-likelihood

$$l_n(\theta) = \sum_{i=1}^{n} \ln[f(x_i;\theta)]$$

since the "ln" is strictly increasing, the MLE of θ can also be obtained by maximizing the log-likelihood function, i.e.

$$\hat{\theta}_n = argmax_{\theta \in \Theta} \;\; l_n(\theta).$$

McGill

Remark

- It is often much easier to work with the log-likelihood

$$l_n(\theta) = \sum_{i=1}^{n} \ln[f(x_i; \theta)]$$

since the "ln" is strictly increasing, the MLE of θ can also be obtained by maximizing the log-likelihood function, i.e.

$$\hat{\theta}_n = argmax_{\theta \in \Theta} \;\; l_n(\theta).$$

Examples

- We will discuss several examples in class.

Summary

From the examples discussed in class, we observed that:

(1) The MLEs are functions of sufficient statistics.

(2) The MLEs are sometime biased, but asymptotically unbiased.

(3) The MLE method (often) yields estimators that are MVUE once the bias is corrected.

MLE and Sufficiency

- Recall the Fisher-Neyman Factorization Theorem, where we have

$$L_n(\theta) = g(t; \theta) \times h(x_1, x_2, \ldots, x_n)$$

and $t = T(x_1, x_2, \ldots, x_n)$.

- The log-likelihood is then given by

$$l_n(\theta) = \ln[g(t; \theta)] + \ln[h(x_1, x_2, \ldots, x_n)].$$

which implies that the MLE of θ is $\hat{\theta}_n = argmax_{\theta \in \Theta} \ln[g(t; \theta)]$.

- Therefore, the MLE of θ is a function of the sufficient statistic $T(X_1, X_2, \ldots, X_n)$.

MLE and Sufficiency

- Recall the Fisher-Neyman Factorization Theorem, where we have

$$L_n(\theta) = g(t;\theta) \times h(x_1, x_2, \ldots, x_n)$$

and $t = T(x_1, x_2, \ldots, x_n)$.

- The log-likelihood is then given by

$$l_n(\theta) = \ln[g(t;\theta)] + \ln[h(x_1, x_2, \ldots, x_n)].$$

which implies that the MLE of θ is $\hat{\theta}_n = argmax_{\theta\in\Theta} \ln[g(t;\theta)]$.

- Therefore, the MLE of θ is a function of the sufficient statistic $T(X_1, X_2, \ldots, X_n)$.

The invariance property of MLE

- Theorem:

  Let $\hat{\theta}_n$ be the MLE of θ. Let $\eta = \tau(\theta)$ be any function of θ. Then, the MLE of η is given by

  $$\hat{\eta}_n = \widehat{\tau(\theta)} = \tau(\hat{\theta}_n).$$

- The proof is posted on myCourses.

Large sample (or asymptotic) properties of the MLE

- Theorem: Under standard REGULARITY CONDITIONS on the family $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, as $n \to \infty$ the MLE $\hat{\theta}_n$ satisfies:

(1) CONSISTENCY: $\hat{\theta}_n \xrightarrow{p} \theta$,

(2) ASYMPTOTIC NORMALITY: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$, where $I(\theta)$ is called the Fisher Information Matrix and is given by

$$I(\theta) = E\left\{\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^\top\right\}$$

which is of dimension $d \times d$.

Large sample (or asymptotic) properties of the MLE

- Theorem: Under standard REGULARITY CONDITIONS on the family $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, as $n \to \infty$ the MLE $\hat{\theta}_n$ satisfies:

(1) CONSISTENCY: $\hat{\theta}_n \xrightarrow{p} \theta$,

(2) ASYMPTOTIC NORMALITY: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$, where $I(\theta)$ is called the Fisher Information Matrix and is given by

$$I(\theta) = E\left\{\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^{\top}\right\}$$

which is of dimension $d \times d$.

Large sample (or asymptotic) properties of the MLE

- Theorem: Under standard REGULARITY CONDITIONS on the family $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, as $n \to \infty$ the MLE $\hat{\theta}_n$ satisfies:

(1) CONSISTENCY: $\hat{\theta}_n \xrightarrow{p} \theta$,

(2) ASYMPTOTIC NORMALITY: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$,
where $I(\theta)$ is called the Fisher Information Matrix and is given by

$$I(\theta) = E\left\{ \left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right] \left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^\top \right\}$$

which is of dimension $d \times d$.

Remarks

- Under the REGULARITY CONDITIONS,

$$I(\theta) = E\left\{ \left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^{\top} \right\} = -E\left\{ \frac{\partial^2 \ln f(X;\theta)}{\partial \theta \partial \theta^{\top}} \right\}.$$

- Intuitively, the Fisher Information matrix captures the variability of the gradient function $\frac{\partial \ln f(X;\theta)}{\partial \theta}$.

- In a parametric family $\mathcal{F}$, for which the gradient has higher variation, intuitively we would except the estimation of θ based on $l_n(\theta)$ be easier; different values of θ change the behaviour of $\frac{\partial \ln f(X;\theta)}{\partial \theta}$ though the log-likelihood function $l_n(\theta)$ varies more.

Remarks

- Under the REGULARITY CONDITIONS,

$$I(\theta) = E\left\{ \left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^\top \right\} = -E\left\{ \frac{\partial^2 \ln f(X;\theta)}{\partial\theta\partial\theta^\top} \right\}.$$

- Intuitively, the Fisher Information matrix captures the variability of the gradient function $\frac{\partial \ln f(X;\theta)}{\partial\theta}$.

- In a parametric family $\mathcal{F}$, for which the gradient has higher variation, intuitively we would except the estimation of θ based on $l_n(\theta)$ be easier; different values of θ change the behaviour of $\frac{\partial \ln f(X;\theta)}{\partial\theta}$ though the log-likelihood function $l_n(\theta)$ varies more.

Remarks

- Under the REGULARITY CONDITIONS,

$$I(\theta) = E\left\{\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^\top\right\} = -E\left\{\frac{\partial^2 \ln f(X;\theta)}{\partial\theta\partial\theta^\top}\right\}.$$

- Intuitively, the Fisher Information matrix captures the variability of the gradient function $\frac{\partial \ln f(X;\theta)}{\partial\theta}$.
- In a parametric family $\mathcal{F}$, for which the gradient has higher variation, intuitively we would except the estimation of θ based on $l_n(\theta)$ be easier; different values of θ change the behaviour of $\frac{\partial \ln f(X;\theta)}{\partial\theta}$ though the log-likelihood function $l_n(\theta)$ varies more.

MLE and Efficiency

- Cramér-Rao inequality: For any unbiased estimator $\tilde{\theta}_n$ of θ, under certain regularity conditions, we have that

$$Var(\tilde{\theta}_n) \geq [nI(\theta)]^{-1}.$$

- This means the MLE is asymptotically (Fisher) efficient ! i.e., it has the smallest possible variance asymptotically.

MLE and Efficiency

- Cramér-Rao inequality: For any unbiased estimator $\tilde{\theta}_n$ of θ, under certain regularity conditions, we have that

$$Var(\tilde{\theta}_n) \geq [nI(\theta)]^{-1}.$$

- This means the MLE is asymptotically (Fisher) efficient !
i.e., it has the smallest possible variance asymptotically.

Note on the regularity conditions

These conditions hold in most cases. However, care must be taken when:

(1) the true value of θ lies on the boundary of the parameter space;

(Example: mixture models)

(2) the support of $f(.;\theta)$ depends on θ.

(Example: $X \sim Unif(0, \theta)$)

Numerical computations of MLE

- MLEs are available in closed form in some parametric families only.
- Typically, numerical optimization methods must be used to obtain MLEs.
- If the log-likelihood is convex and smooth, numerical methods work well!
- Moment estimates provide good starting values which are essential in most of the optimization methods.

MLE in R

- MLE is implemented in R for many univariate distributions such as:

  Beta, Cauchy, Chi-squared, Exponential, F, Gamma, Geometric, Log-normal, Lognormal, Logistic, Negative binomial, Normal, Poisson, t, Weibull.

- Using R, we will discuss the birth time example in class.

MLE in R

- MLE is implemented in R for many univariate distributions such as:

  Beta, Cauchy, Chi-squared, Exponential, F, Gamma, Geometric, Log-normal, Lognormal, Logistic, Negative binomial, Normal, Poisson, t, Weibull.

- Using R, we will discuss the birth time example in class.