Probability and Statistics

Omid Safarzadeh

December 20, 2021

Table of contents

*Acknowledgement: This slide is prepared based on Casella and Berger, 2021

Introduction

In this part, we will talk about estimation. Our focus will almost exclusively be on the maximum likelihood method. We have worked with many distributions so far, calculated their expectations, variances or derived their moment generating functions etc. Importantly, the setting was such that we knew what distribution we were considering AND we had full knowledge of the parameter values for these distributions. Or, to put it more precisely, we never contemplated the possibility that they might not be known. Then, there are two implicit assumptions:

1. We know the distribution
2. We know the parameters of the distribution.

In real life, this is rarely the case. We will first relax the second assumption and later on will dispense with the first assumption. The treatment will sacrifice on formality and will rather focus on ideas. References for more formal treatments will be provided at the end of this set of slides.

Introduction

Now, let's assume we have a random sample consisting of $X_1, ..., X_n$ from the density $f_X(x|\theta_0)$. We would like to determine the value of θ_0, which is unknown. We could use an estimator. An estimator is some function of the data

$$\hat{\theta_n} = W(X_1, ..., X_n). \quad (1)$$

The index n underlines that fact that the particular value of the estimate depends on the sample (and, so, on its size). Note that usually n is dropped and instead simply $\hat{\theta}$ is used. Note the difference between the estimator and the estimate.The estimator is a concept while the estimate is the value of the estimator for a given sample. So, if the estimator is $W(X_1, ..., X_n)$, then the estimate for a particular realisation of $X_1, ..., X_n$ is given by $W(x_1, ..., x_n)$.

Introduction

Now, although the definition given in (1) implies that any function of the data could be a valid estimator, we usually look for those that have desirable properties.In other words, an estimator is a statistic (meaning that it cannot depend on θ or any other unknown parameters) which has desirable properties. We have actually introduced one of these desirable properties: consistency.Others are unbiasedness, minimum mean squared error, minimum variance etc. Let Θ be the parameter space for θ. An estimator $\hat{\theta}$ of θ_0 is a minimum mean squared error estimator if for every $\theta_0 \in \Theta$.

$$\hat{\theta} = arg \min_{\theta \in \Theta} E[(\theta - \theta_0)^2].$$

An estimator $\hat{\theta}$ of θ_0 is unbiased if for every $\theta_0 \in \Theta$,

$$E[\hat{\theta}] = \theta_0.$$

You will learn more about these in your future econometrics courses

Motivation and the Main Ideas

Let us first disect the notation. Suppose we are dealing with some generic distribution such that

$$F_Y(y;\theta), \quad \theta \in \Theta.$$

F is the cdf, Y is the random variable and y is a particular realisation of Y. θ is a vector which contains the distribution parameters. This is generally known as the parameter vector.The parameter vector takes on values on a set, Θ, known as the parameter space.For example, for a normal random variable,

$$\theta = (\mu, \sigma^2) \quad \text{and} \quad \Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\},$$

where μ is the mean and σ^2 is the variance.

Motivation and the Main Ideas

Suppose that we actually know the distribution. However, usually θ_0 is unknown. How to find out the value of θ_0? We have to distinguish between the population and the sample. Population contains all the unknown values. The sample, on the other hand can only provide an approximation. For example,

$$\theta_0 \ : \ \text{population}, \hat{\theta} \ : \ \text{sample}$$

The maximum likelihood method is a very popular and strong method for estimating θ when the underlying distribution function, F_Y, is known (or when one believes that one actually knows the underlying distribution).

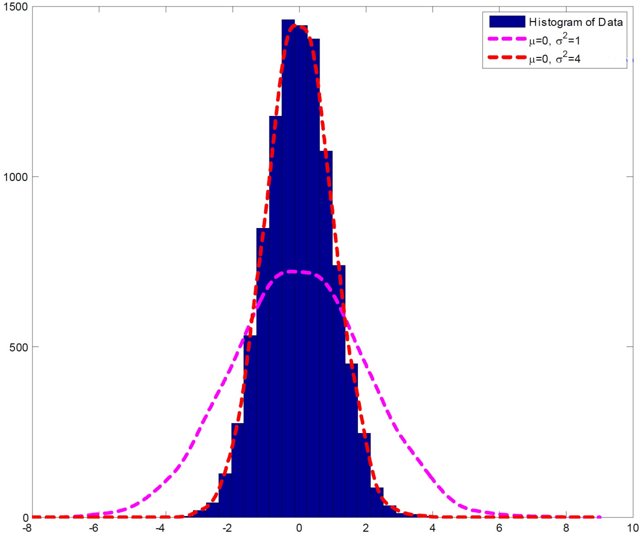
Motivation and the Main Ideas

Maximum likelihood estimation (MLE) is based on maximisation of a likelihood function.Where to find this "likelihood function?" It's actually pretty easy! The likelihood function is the same as the probability density function:

$$f_Y(y;\theta) = L(\theta : y).$$

The only change is the interpretation. When we consider a probability density function, we implicitly consider θ as fixed and y as random. When we consider a likelihood function, we assume that data, y, are given and fixed. Instead, it is θ which is modified.How to make sense of this? MLE is based on the idea that, if we know the underlying distribution function, then we should choose θ such that the probability of the data, y, being observed is maximised. In other words, we are trying to find out the values of θ which are most likely to generate the observed data. This likelihood principle is due to R. A. Fisher.

Motivation and the Main Ideas



Motivation and the Main Ideas

The dataset would preferably consist of many observations on the same random variable. This ensures that we have sufficient information to estimate θ. Consider some simple examples.

Example 2.1

Let Y_i be an iid random sequence where $i = 1, ..., n$. Let also $Y_i \sim N(\mu, \sigma^2)$, where $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$ gives the parameter space. Then, thanks to the independence assumption, the joint likelihood function is given by

$$f_Y(y; \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta),$$

where $y = (y_1, ..., y_n)$.

- Notice that the parameter vector, q, is common for all variables.

Example 2.2

Let Y_i be an iid random sequence, conditional on $Xi = xi$, where $i = 1, ..., n$. Let also $Y_i|X_i = x_i \sim N(\beta' x_i, \sigma^2)$, where $x_i = (x_{i1}, ..., x_{ik})'$ and $\beta = (\beta_1, ..., \beta_k)'$. Let, also

$$y = (y_1, ..., y_n) \quad \text{and} \quad x = (x_1, ..., X_n).$$

Then,

$$f_{Y|X=x}(y;\theta) = \prod_{i=1}^{n} f_{Y_i|X_i=x_i}(y_i;\theta).$$

Motivation and the Main Ideas

Example 2.3

A possible structure for $Y_i|X_i = x_i$ would be

$$yi = 0.3x_{i1} + 0.4x_{i2} + u_i, \quad u_i \overset{iid}{\sim} N(0,1),$$

Then,

$$f_{Y|X=x}(y;\theta) = \prod_{i=1}^{n} f_{Y_i|X_i=x_i}(y_i;\theta).$$

and x_{i1} and x_{i2} have their own distributions, where u_i is independent of $x_i = (x_{i1}, x_{i2})$. Here,

$$\beta = (0.3, 0.4)', \ \ \sigma^2 = 1 \ \text{ and } \ x_i = (x_{i1}, x_{i2})'.$$

Maximum Likelihood Estimation

One of the most important models in financial econometrics (and, indeed, econometrics) is the autoregressive conditional heteroskedasticity (ARCH) mode due to Engle (1982, Econometrica).

Example 2.4

Let Y_t be the daily return on some equity on day t, where $t = 1, ..., T$. The model is given by

$$Y_t | Y_{t-1} = y_{t-1} \sim N(0, \sigma_t^2),$$

where

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2.$$

Maximum Likelihood Estimation

We can construct the likelihood by using a representation known as prediction decomposition. Omitting the arguments of the likelihood/density function for conciseness, we obtain

$$f_{Y_1,\ldots,Y_T} = f_{Y_2,\ldots,Y_T|Y_1} f_{Y_1}$$

$$= f_{Y_3,\ldots,Y_T|Y_1,Y_2} f_{Y_2|Y_1} f_{Y_1}$$

$$= f_{Y_4,\ldots,Y_T|Y_1,Y_2,Y_3} f_{Y_3|Y_1,Y_2} f_{Y_2|Y_1} f_{Y_1}$$

$$.$$

$$.$$

$$.$$

$$= f_{Y_1} \prod_{t=2}^{T} f_{Y_t|Y_{t-1},\ldots,Y_1}$$

Maximum Likelihood Estimation

Then, for the ARCH model we have

$$f_{Y_1,\ldots,Y_T} = f_{Y_1} \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}$$

$$\approx \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}$$

where we use the information that the conditional distribution of Y_t depends on Y_{t-1} only.

Maximum Likelihood Estimation

Now that we know how to construct the joint likelihood function for a collection of random variables $Y_1, ..., Y_n$, we can start thinking about how to estimate parameters by MLE. Remember our discussion about the logic behind MLE. The idea is to find the values of the parameters that maximise the possibility of obtaining the data that we observe in the sample. Our notation is

$$L(\theta; y) = f_Y(y; \theta),$$

where θ and y are the parameter and data **matrices**, respectively. Usually, it is more convenient to use the log-likelihood which is

$$\ell(\theta; y) = \log L(\theta; y).$$

Notice that log is a monotone transformation. Hence, as will be obvious in a moment, for our purposes there is no difference between using $\ell(\theta; y)$ and $L(\theta; y)$. The maximum likelihood method is based on finding the parameter values which maximise the likelihood (or probability) of obtaining the particular sample we have:

$$\hat{\theta} = arg \max_{\theta \in \Theta} \ell(\theta; y).$$

Maximum Likelihood Estimation

Hence, the likelihood function is the objective function. Consequently, there must be a first order condition.

- **Caution**: never confuse estimator with estimate!

This first order condition has a special name: the score. The score is a key concept and deeply influences the behaviour of the ML estimator. Let θ be a $(k * 1)$ vector. When the derivative exists, the score is given by

$$\frac{\partial \log L(\theta; y)}{\partial \theta} = \frac{\partial \ell(\theta; y)}{\partial \theta}.$$

Of course, this is a $(k * 1)$ vector, as well. Consequently, $\hat{\theta}$ is the value of θ which satisfies,

$$\frac{\partial \log L(\theta; y)}{\partial \theta}|_{\theta=\hat{\theta}} = 0.$$

Importantly, one also has to ensure that

$$\frac{\partial^2 \log L(\theta; y)}{\partial \theta \partial \theta'}|_{\theta=\hat{\theta}} < 0,$$

in the sense that the matrix is negative definite.

Maximum Likelihood Estimation

Example 2.5

Let $Y \sim N(\mu, \sigma^2)$ where $Y_i \perp\!\!\!\perp Y_j$ for all $i \neq j$ and $i, j = 1, ..., n$. Let, as before, $y = (y_1, ..., y_n)$. The joint likelihood is given by

$$L(\theta; y) = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta)$$

$$(\frac{1}{\sqrt{2\pi\sigma^2}})^n exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\}.$$

Then,

$$l(\theta; y) = \log L(\theta; y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Obviously, $\theta = (\mu, \sigma^2)$. Let's find the ML estimators.

Maximum Likelihood Estimation

Now,

$$\frac{\partial \ell(\theta; y)}{\partial \mu}\Big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma^2}} = \frac{1}{\hat{\sigma^2}} \sum_{i=1}^{n} (y_i - \hat{\mu}) = 0,$$

and

$$\frac{\partial \ell(\theta; y)}{\partial \mu}\Big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma^2}} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = 0.$$

Solving the first-order conditions for the parameters yields,

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \text{ and } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Therefore, $\hat{\theta} = (\bar{y}, \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2)'$.

Properties of the Maximum Likelihood Estimator

In the next few slides, we will cover some important common properties of likelihood functions. In this discussion, we will assume that the data generating process and the chosen underlying distribution are the same:

$$g_Y(y) = f_Y(y; \theta_0).$$

where $g_Y(y)$ is the **true** data generating process and θ_0 is, by definition, the true parameter value. This is not necessarily true in general. In fact, thinking about what happens when

$$g_Y(y) \neq f_Y(y; \theta) \quad \text{for all possible } \theta$$

is crucial. We will do this later. For the time being, we will stick to the simpler case.

Properties of the Maximum Likelihood Estimator

In what follows, it will be important to make it clear according to which density we are taking the expectation or the variance.In general, for any function $A(X)$ where X is some random variable, we will use

$$E_f[A(X)] = \int A(X) f(x;\theta_0) dx$$

$$Var_f(A(X)) = \int \{A(X) - \mu_A\}^2 f(x;\theta_0) dx,$$

where $\mu_A = E_f[A(X)]$, etc. Sometimes, we will also make this more explicity by using the index $f(x|\theta_0)$, e.g.,

$$E_{f(x|\theta_0)}[A(X)] = \int A(x) f(x;\theta_0) dx.$$

We do not necessarily have to take all the moments with respect to the true density function (or, the data generating process), of course. For example, one might also be interested in

$$E_{g(x|\psi)}[A(X)] = \int A(x) g(x;\psi) dx,$$

where $g(x;\psi)$ is another distribution for X with the parameter ψ.

- **Property 1 (Unbiasedness of the Score):**

$$E_f[\frac{\partial \log L(\theta, Y)}{\partial \theta}|_{\theta=\theta)}] = 0,$$

where the expectation is taken with respect to the distribution $f_Y(y; \theta_0)$.

- **Proof**: Now,

$$\frac{\partial \log L(\theta, Y)}{\partial \theta} = \frac{1}{L(\theta, Y)} \frac{\partial L(\theta; y)}{\partial \theta}.$$

Then,

$$E_f[\frac{\partial \log L(\theta, Y)}{\partial \theta}] = \int \{\frac{1}{L(\theta, Y)} \frac{\partial L(\theta; y)}{\partial \theta}\} f(y; \theta_0) dy,$$

by definition. Observe that this is a function of both θ and θ_0.

Properties of the Maximum Likelihood Estimator

Now,

$$E_f[\frac{\partial \log L(\theta, Y)}{\partial \theta}|_{\theta=\theta_)}] = \int [\frac{\partial L(\theta; y)}{\partial \theta} \partial 1 L(\theta; y)]|_{\theta=\theta_0} f(y; \theta_0) dy$$

$$\int \frac{\partial L(\theta; y)}{\partial \theta}|_{\theta=\theta_0} dy$$

$$= (\frac{\partial}{\partial \theta} \int f(y; \theta) dy|_{\theta=\theta_0}$$

$$= (\frac{\partial}{\partial \theta} 1)|_{\theta=\theta_0}$$

$$= 0,$$

where we implicitly assumed that the order of integration and differentiation can be exchanged. *An* aside: this requires that the range of y does not depend on θ. Hence, the expectation of the first-order condition, evaluated at the true parameter value, is zero!

Properties of the Maximum Likelihood Estimator

- **Property 2 (The Information Equality)**:

$$Cov_f(\frac{\partial \log L(\theta; Y)}{\partial \theta}|_{\theta=\theta_0}) = -E_f[\frac{\partial^2 \log L(\theta; Y)}{\partial \theta \partial \theta'}|_{\theta=\theta_)}],$$

  where, as before, the expectation and the covariance are taken with respect to $f_Y(y; \theta)$.

Properties of the Maximum Likelihood Estimator

- **Property 3 (Cramér-Rao Inequality)**: Let $\theta^{\sim}$ be some estimator of $\theta_{)}$ and assume that $E_f[\theta^{\sim}] = \theta_0$. Then,

$$Var_f(\widetilde{\theta}) - \{Var_f[\frac{\partial \log L(\theta; Y)}{\partial \theta}|_{\theta=\theta_0}]\}^{-1} \geq 0,$$

in the sense that the difference between the two matrices is non-negative definite.

Properties of the Maximum Likelihood Estimator

We note two important results, without proving them:

1. The only time the Cramér-Rao bound is achieved is when the estimator is the ML estimator. In many problems, no estimator would actually achieve this bound.
2. For regular problems, asymptotically the ML estimator achieves the Cramér-Rao bound. In other words, for large n, ML can achieve the Cramér-Rao bound, that is ML is efficient in large samples.

Properties of the Maximum Likelihood Estimator

Now, the previous term looks like the covariance of

Example 2.6

Let $Y_1, Y_2, ...$ be an iid sequence where $Y_i \sim N(\theta_0, 1)$ for all i. We will first find the Cramér-Rao bound for unbiased estimators of q0 and then show that the ML estimator achieves this bound. First, let's construct the log-likelihood function. Let $y = (y_1, ..., y_n)'$. Then

$$L(\theta; y) = \frac{1}{\sqrt{2\pi 1}} exp\{-\frac{1}{2}\frac{(y_1-\theta)^2}{1}\} \cdots\cdots \frac{1}{\sqrt{2\pi 1}} exp\{-\frac{1}{2}\frac{(y_n-\theta)^2}{1}\}$$

$$= (\frac{1}{\sqrt{2\pi}})^n exp\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2\}.$$

This gives

$$\ell(\theta; y) = -\frac{n}{2}\log 2\pi - -\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2.$$

Now, the previous term looks like the covariance of

Example 2.7

Then, the first order condition is given by

$$\frac{\partial \ell(\theta; y)}{\partial \theta}|_{\theta=\theta_0} = -\frac{1}{2}(-2)\sum_{i=1}^{n}(y_i - \hat{\theta}) = 0,$$

implying that

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\theta} = 0,$$

and, so,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Example 2.7 cont.

In addition,

$$Var_f(\frac{\partial \log f(y;\theta)}{\partial \theta}|_{\theta=\theta_0}) = Var_f[\sum_{i=1}^{n}(y_i - \theta_0)]$$

$$= \sum_{i=1}^{n} \underbrace{Var_f(y_i)}_{1} = n,$$

due to the iid assumption.

Properties of the Maximum Likelihood Estimator

Therefore, as far as this problem is concerned, the Cramér-Rao bound for any unbiased estimator $\theta^{\sim}$ is given by

$$Var_f(\widetilde{\theta}) \geq \frac{1}{n}.$$

Now, the variance of the ML estimator is very easy to find.

$$Var_f(\widetilde{\theta}) = Var(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(y_i)$$

$$= \frac{1}{n^2} n = \frac{1}{n}.$$

But this the same as the Cramér-Rao bound. Hence, the ML estimator in this particular case is an efficient estimator.

Properties of the Maximum Likelihood Estimator

Remember that we restrict ourselves to the case where our random sequence $Y_1, ..., Y_n$ is iid. Let $y = (y_1, ..., y_n)$. Let $f_{Y_i}(y_i; \theta) = L(\theta; y_i)$ be the pdf (or the likelihood function) for Y_i. Then, the joint pdf or the joint likelihood function is given by

$$L(\theta; y) = \prod_{i=1}^{n} L(\theta; y_i),$$

and the joint log-likelihood function is

$$\ell(\theta; y) = \log \prod_{i=1}^{n} L(\theta; y_i) = \sum_{i=1}^{n} \log L(\theta; y_i).$$

Now, suppose that $\log L(\theta; y_1), \log L(\theta; y_2), ...$ is an iid sequence where $E[|\log L(\theta; y_i)|] < \infty$ for all i. Then, by the strong Law of Large Numbers,

$$\frac{1}{n} \sum_{i=1}^{n} \log L(\theta; y_i) \xrightarrow{a.s.} E_{f(y|\theta_0)}[\log L(\theta; Y_i)].$$

References

Casella, G., & Berger, R. (2021). *Statistical inference*. Cengage Learning. https://books.google.fr/books?id=FAUVEAAAQBAJ