Estimators

MFM Practitioner Module: Risk & Asset Allocation

John Dodson

October 5, 2011
Outline

Motivation

Sample
  Sufficient Statistic

Estimator
  Loss

Maximum Likelihood Estimator
  Standard Error

Admissibility
The Goal of Estimation

The goal of estimation is to assign numerical values to the parameters of a probability model.

Considerations
There are several risks to consider:

▶ What if the model is mis-specified?
▶ What if the data are corrupt?

These are addressed under the subject of robust statistics, which we will cover in the last week this term.
In classical statistics, the term *sample* has two related meanings

- an (unordered) set of $N$ values drawn from the sample space of some random variable $X$, $\{x_1, x_2, \ldots, x_N\}$
- a random variable consisting of $N$ (independent) copies $X_1, \ldots, X_N$ of some random variable $X_i \sim X \ \forall i$.

You can think of the former as a realization of the latter. We can characterize the latter, which we will denote hereafter by $Y^{(N)} = (X_1, \ldots, X_N)$, as a random variable with

$$f_{Y^{(N)}}(Y) = f_X(X_1) \cdots f_X(X_N)$$

because we have assumed that the draws are independent.
Sufficient Statistic

The characterization of the sample $Y^{(N)}$ can often be expressed as the characterization of a collection of partial results, $T = T(X_1, \ldots, X_N; N)$, called sufficient statistics.

Important Example

Say $X \sim \mathcal{N}(\mu, \sigma^2)$ and we have a sample $Y^{(N)} = (X_1, \ldots, X_N)$. The density function of the sample is

$$f_{Y^{(N)}}(y) = (2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{N}(x_i - \mu)^2}$$

The form of this suggests $T = (\sum X_i, \sum X_i^2; N)$, which yields

$$f_T(t) = \frac{(Nt_2 - t_1^2)^{(N-3)/2}}{N^{N/2-1}2^{N/2}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N-1}{2}\right)}$$

$$\cdot \exp\left(\frac{1}{\sigma^2} \left( \frac{t_2}{N} - 2 \frac{t_1}{N} \mu + \mu^2 \right) + \log \sigma^2\right)^{-N/2}$$

\[\ast\]
Classical Estimator

An estimator is a function of a sample.

- If the sample is considered to be random, the value of an estimator is a random variable subject to characterization.

- If the estimator is applied to an actual sample, consisting of $N$ draws from the sample space, the value is non-random and is called an estimate.

Parameter Estimator

We will be mostly interested in estimating the parameters of a characterization, which we will denote generically by $\theta$. For a univariate normal, for example, $\theta = (\mu, \sigma^2)'$.

We will denote the parameter estimator by $\hat{\theta}(Y^{(N)})$ where $Y^{(N)} = (X_1, \ldots, X_N)$ is the sample represented by $N$ independent copies of the random variable $X$ with a characterization parameterized by $\theta$. 
Quadratic Loss

Since \( \hat{\theta}(Y^{(N)}) \) is a random variable, it is natural to explore its location and dispersion.

- In particular, we are interested in how far it can diverge from the (unknown) true value, \( \theta \).
- So we introduce a norm with respect to some positive definite metric \( Q \), such that \( \|v\|^2 = v'Qv \) for any \( v \) in the sample space of \( \theta \).
- **Loss** is the random variable \( \|\hat{\theta} - \theta\|^2 \).
- **Bias** is the (unknown) value \( \|E\hat{\theta} - \theta\| \).
- **Inefficiency** is the value \( \sqrt{E\|\hat{\theta} - E\hat{\theta}\|^2} \).

There is a trade-off between bias and inefficiency. In fact,

\[
E \text{ Loss} = \text{Bias}^2 + \text{Inef}^2 \quad \text{(prove)}
\]
Maximum Likelihood Estimator

Since we have the distribution of the sample, perhaps in terms of sufficient statistics, it is natural to define an estimator for the parameters as the value of the parameters such that the sample observed is “most likely”. That is,

$$\hat{\theta}(y) = \arg \max_{\theta} f_{Y(N)}|_{\theta}(y) \quad \text{or} \quad \hat{\theta}(y) = \arg \max_{\theta} f_{T}|_{\theta}(t)$$

where the sample is $y = (x_1, \ldots, x_N)$ or $t = T(x_1, \ldots, x_N; N)$.

Important Example

Consider the univariate normal from above. In terms of the sufficient statistics, the MLE (based on $(\ast)$) is

$$\left( \begin{array}{c} \hat{\mu} \\ \hat{\sigma}^2 \end{array} \right) = \arg \min_{(\mu, \sigma^2)} \frac{1}{\sigma^2} \left( \frac{t_2}{N} - 2 \frac{t_1}{N} \mu + \mu^2 \right) + \log \sigma^2$$
Important Example

The solution to this (the MLE for a univariate normal) is

\[ \hat{\mu} = \frac{t_1}{N} \]
\[ \hat{\sigma}^2 = \frac{t_2}{N} - \left( \frac{t_1}{N} \right)^2 = \frac{xx'}{1'1} - \frac{1'x'x1}{1'1'11} \]

This result extends to the mutivariate case \( X \in \mathbb{R}^M \) whereby \( x \) has \( M \) rows and \( N \) columns.

Bias

We can see that the MLE is (slightly) biased.

\[ E \hat{\mu} = \mu \]
\[ E \hat{\sigma}^2 = \frac{N - 1}{N} \sigma^2 \quad \text{(prove)} \]
Maximum Likelihood Estimator

Elliptical random variables

If the density of an r.v. $X \in \mathbb{R}^M$ can be written in the form

$$f_{X|\mu,\Sigma}(x) = g \left( \text{Ma}^2(x, \mu, \Sigma) \right) \sqrt{|\Sigma^{-1}|}$$

for some function $g(\cdot)$, then the MLE based on a sample \{\(x_1, \ldots, x_N\}\) solves the system

$$\hat{\mu} = \sum_{i=1}^{N} \frac{w_i}{\sum_j w_j} x_i$$

$$\hat{\Sigma} = \sum_{i=1}^{N} \frac{w_i}{N} (x_i - \hat{\mu}) (x_i - \hat{\mu})'$$

with

$$w_i = \frac{-2g' \left( \text{Ma}^2 \left( x_i, \hat{\mu}, \hat{\Sigma} \right) \right)}{g \left( \text{Ma}^2 \left( x_i, \hat{\mu}, \hat{\Sigma} \right) \right)} \quad \forall i = 1, \ldots, N$$
Fisher Information

In general we cannot evaluate the characterization of the distribution of an estimator. An application of the Central Limit Theorem gives us a useful approximation.

\[
\lim_{N \to \infty} \sqrt{N} \left( \hat{\theta} \left( Y^{(N)} \right) - \theta \right) \sim \mathcal{N} \left( 0, I_{X|\theta}^{-1} \right)
\]

where \( I \) is the Fisher Information matrix

\[
I_{X|\theta} = \text{cov} \left( \frac{\partial}{\partial \theta^i} \log f_{X|\theta}(X) \right) = -\mathbb{E} \frac{\partial^2}{\partial \theta \partial \theta^i} \log f_{X|\theta}(X)
\]

Important Example

For the univariate normal, this evaluates to

\[
I_{X|\mu,\sigma^2} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}
\]
Cramér-Rao Bound

The Cramér-Rao Bound gives us a limit on the resolution of an estimator.

\[
\text{cov} \left( \hat{\theta} \left( Y^{(N)} \right) \right) \geq \frac{\partial E \hat{\theta}}{\partial \theta} \frac{I^{-1}}{N} \frac{\partial E \hat{\theta}'}{\partial \theta'}
\]

which is attained if the estimator is efficient.

Standard Error

The standard deviations of the margins of the estimator are called the standard errors

\[
\text{se}(\hat{\theta}) = \text{diag} \sqrt{\text{diag diag cov } \hat{\theta}}
\]

In the case of the univariate normal example, the bound is

\[
\text{se} \left( \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \right) \geq \begin{pmatrix} \sigma \sqrt{N} \\ \sigma^2 \sqrt{N/2} \frac{N-1}{N} \end{pmatrix}
\]
Admissibility

The expected value of an estimator’s loss (given the unknown true value) is also called the mean squared error\(^1\). We saw before that this is the sum of the estimator’s squared bias and squared inefficiency.

Admissible Estimators
An estimator whose expected loss is no greater than that of any other estimator for all possibilities of the unknown value is termed admissible.

Inadmissibility of the sample mean
We know from the Law of Large Numbers that the sample mean estimator is unbiased and is efficient in the limit of large samples. It should come as shock then that, with a sample space of at least three dimensions, the sample mean is inadmissible.

\(^1\)Elsewhere this is called the estimator’s risk.
Shrinkage Estimator

It turns out that for $X \in \mathbb{R}^M$ with $M > 2$ and a sample of length $N$, an estimator based on shrinking the sample mean towards any arbitrary value $\mu_0 \in \mathbb{R}^M$ by a particular amount $0 < \alpha_0 < 1$ has lower expected loss.

$$\hat{\mu} = \alpha_0 \mu_0 + (1 - \alpha_0) \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_0)$$

For the optimal $\alpha_0$, this is termed the James-Stein estimator.

- Of course, unless $\mu_0$ happens to equal the true value, this estimator is biased.
- But the reduced inefficiency makes the bias worth it.
- Yet the result is still inadmissible. Improving upon it is still an open question in statistics.

Leading Question

Where does the value $\mu_0$ come from?