Estimators

MFM Practitioner Module: Quantitiative Risk Management

John Dodson

September 30, 2015

Outline

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 briefly introduce.

Sample

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 $Y^{(N)} \triangleq (X_1, \ldots, X_N)$, 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

$$f_T(t) = \frac{\left(Nt_2 - t_1^2\right)^{(N-3)/2}}{N^{N/2-1}2^{N/2}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N-1}{2}\right)} \cdot \exp\left(\frac{1}{\sigma^2}\left(\frac{t_2}{N} - 2\frac{t_1}{N}\mu + \mu^2\right) + \log\sigma^2\right)^{-N/2} \quad (*)$$

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 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 θ. For a univariate normal, for example, $\theta = \left(\mu, \sigma^2\right)'$.

We will denote the parameter estimator by $\hat{\theta}\left(Y^{(N)}\right)$ 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 θ.

Quadratic Loss

Since $\hat{\theta}\left(Y^{(N)}\right)$ 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, θ.
- 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 θ.
- Loss is the random variable $\|\hat{\theta} - \theta\|^2$.
- Bias is the (unknown) value $\|\operatorname{E}\hat{\theta} - \theta\|$.
- Inefficiency is the value $\sqrt{\operatorname{E}\|\hat{\theta} - \operatorname{E}\hat{\theta}\|^2}$.

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

$$\operatorname{E}\text{Loss} = \text{Bias}^2 + \text{Inef}^2 \quad \text{(prove)}$$

Method of Moments

One classical method for estimating the parameters of a random variable from a sample is to identify low-order sample moments with the corresponding "population" moments of the random variable.

- sample mean $\bar{x} \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i$
- sample variance $\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$

Moment Matching

If a random variable X has a parametric characterization with only one or two parameters, $\theta = (\theta_1, \theta_2)'$, it is likely that a system of the form

$$\begin{cases} \mathrm{E}\, X|\hat{\theta}\left(Y^{(N)}\right) = \frac{1}{N}\sum_{i=1}^{N} X_i \\ \mathrm{var}\, X|\hat{\theta}\left(Y^{(N)}\right) = \frac{1}{N-1}\sum_{i=1}^{N} X_i^2 - \frac{1}{N(N-1)}\left(\sum_{i=1}^{N} X_i\right)^2 \end{cases}$$

implicitly defines a unique solution for $\hat{\theta}\left(Y^{(N)}\right)$.

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,

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

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 $(*)$) is

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

Maximum Likelihood Estimator

Important Example

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

$$\hat{\mu} = \frac{t_1}{N} = \frac{x1}{1'1}$$

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

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

Bias

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

$$\mathrm{E}\ \hat{\mu} = \mu$$

$$\mathrm{E}\ \hat{\sigma^2} = \frac{N-1}{N}\sigma^2 \quad \text{(prove)}$$

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 $g(\cdot)$ where

$$\text{Ma}(x,\mu,\Sigma) = \sqrt{(x-\mu)'\,\Sigma^{-1}\,(x-\mu)}$$

is the Mahalanobis distance, then the MLE based on a sample $\{x_1,\ldots,x_N\}$ solves the system

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

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

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

$$\begin{aligned} I_{X|\theta} &= \operatorname{cov} \frac{\partial}{\partial \theta'} \log f_{X|\theta}(X) \\ &= -\operatorname{E} \frac{\partial^2}{\partial \theta \, \partial \theta'} \log f_{X|\theta}(X) \end{aligned}$$

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.

$$\operatorname{cov} \hat{\theta}\left(Y^{(N)}\right) \geq \frac{\partial \operatorname{E} \hat{\theta}}{\partial \theta} \frac{I_{X|\theta}^{-1}}{N} \frac{\partial \operatorname{E} \hat{\theta}'}{\partial \theta'}$$

which is attained if the estimator is efficient.

Standard Error

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

$$\operatorname{se}(\hat{\theta}) = \operatorname{diag} \sqrt{\operatorname{diag} \operatorname{diag} \operatorname{cov} \hat{\theta}}$$

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

$$\operatorname{se} \begin{pmatrix} \hat{\mu} \\ \hat{\sigma^2} \end{pmatrix} \geq \begin{pmatrix} \frac{\sigma}{\sqrt{N}} \\ \frac{\sigma^2}{\sqrt{N/2}} \frac{N-1}{N} \end{pmatrix}$$

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 α_0, this is termed the James-Stein estimator.

- Of course, unless μ_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.

Robustness

Non-Parametric Estimators

The term robustness in statistics can sometimes refer to non-parametric techniques that do not require assumptions about the characterization of the random variables involved.

- Such techniques usually lean on the Law of Large Numbers, and hence require very large samples to be effective.

Robust Estimators

A more precise meaning has evolved that focuses on estimators that may be based on parametric characterizations, but which can produce reasonable results for data that does not come from that class of characterizations or stress-test distributions.

- We can make this desire concrete in term of the the influence function associated with an estimator.

Robust Estimators

Influence Function

Wwe have discussed estimators as functions of samples. If instead we consider the estimator as a functional of the density from which the sample is drawn, we can consider its (functional) derivative with respect to an infinitesimal perturbation in the density given by

$$f_X(x) \to (1-\epsilon)f_X(x) + \epsilon\delta(x-y)$$

Thus, with $\tilde{\theta}$ the functional induced by the estimator $\hat{\theta}$,

$$\text{IF}\left[y, f_X, \hat{\theta}\right] = \lim_{\epsilon \to 0} \frac{\tilde{\theta}\left[(1-\epsilon)f_X(x) + \epsilon\delta(x-y)\right] - \tilde{\theta}\left[f_X\right]}{\epsilon}$$

If this derivative is bounded for all possible displacements, y, we say the estimator is robust.

Robust Estimators

Robustness of the MLE

For the maximum likelihood estimator, the influence function turns out to be proportional to

$$\text{IF}\left[y, f_X, \hat{\theta}\right] \propto \left.\frac{\partial \log f_{X|\theta}(y)}{\partial \theta}\right|_{\theta=\hat{\theta}}$$

For some characterizations, the parameter MLE's are robust. For some they are not.

- for $X \sim \mathcal{N}(\mu, \Sigma)$, $\hat{\mu}$ and $\hat{\Sigma}$ are not robust
- for $X \sim \text{Cauchy}(\mu, \Sigma)$, they are

Even for the empirical characterization, the influence functions for the sample mean and the sample covariance are not bounded; therefore these sample estimators are *never* robust.

M-Estimators

Location and Dispersion

Recall the general elliptic location and dispersion MLE's,

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

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

$$w_i \triangleq h\left(\mathrm{Ma}^2\left(x_i, \hat{\mu}, \hat{\Sigma}\right)\right) \quad \forall i = 1, \ldots, N$$

where the function $h(\cdot)$ is the value of a particular functional on the density. The idea with M-estimators is to choose $h(\cdot)$ exogenously in order to bound the influence function by design.

M-Estimators

We know that $h(\cdot) = 1$ corresponds to the MLE for normals and also to the sample estimators, which do not have bounded influence functions. A weighting function that falls towards zero for large arguments is more likely to be robust. Some examples include

- Trimmed estimators for which
$$h(z) = \begin{cases} 1 & z < z_0 = Q_{\chi^2_{\dim X}}(p) \\ 0 & \text{otherwise} \end{cases}$$
- Cauchy estimators for which $h(z) = \frac{1+\dim X}{1+z}$
- Schemes such as Huber's or Hampel's for which
$$h(z) = \begin{cases} 1 & z < z_0 = \left(\sqrt{2} + \sqrt{\dim X}\right)^2 \\ \sqrt{\frac{z_0}{z}} e^{-\frac{(\sqrt{z}-\sqrt{z_0})^2}{2b^2}} & \text{otherwise} \end{cases}$$

These estimators can be evaluated numerically by iterating to the fixed point.