Estimation: Review

Statistical Models

Data: A random variable (or vector) $X$ has (joint) distribution

The Unknowns: The joint distribution is not completely known.

Estimators: Let $\theta$ be an unknown parameter. An estimator is a function of the data $\hat{\theta}(X)$.

Methods
- The Method of Moments
- Maximum Likelihood

Example: Exponential

Cover Story: $n$ manufactured items are place on test and observed to fail after $X_1, \ldots, X_n$ time units. Suppose

$$X_1, \ldots, X_n \sim \text{exp} \left( \lambda \right)$$

where $\lambda$ is unknown (the failure rate).

The Method of Moments. For and exponential distribution, $f(x) = \lambda e^{-\lambda x}, \ x > 0$, the mean is $\mu = E(X_i) = 1/\lambda$. The MOM solves the equation

$$\mu = \bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

So,

$$\hat{\lambda} = \frac{1}{\bar{X}} = \frac{n}{X_1 + \cdots + X_n}$$

Maximum Likelihood. The likelihood function and log-likelihood function are

$$L(\lambda|x) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(x_1 + \cdots + x_n)}$$

and

$$l(\lambda|x) := \log[L(\lambda|x)] = n \log(\lambda) - \lambda(x_1 + \cdots + x_n).$$

Here

$$\ell'(\lambda|x) = \frac{d}{d\lambda} \ell(\lambda|x) = \frac{n}{\lambda} - (x_1 + \cdots + x_n) = 0$$

iff

$$\hat{\lambda} = \frac{n}{x_1 + \cdots + x_n}.$$

Note: The MLE and MOM are the same in this example.

Properties of Estimators

Notation: Let

$$T = T(X_1, \ldots, X_n)$$

be an estimator of a parameter $\theta$—for example, the sample and population means.

Bias: The bias of $T$ is

$$b_T = E(T) - \theta;$$

and $T$ is said to be unbiased if $b_T \equiv 0$.

Example. The sample mean and variance are unbiased, since $E(\bar{X}) = \mu$ and $E(S^2) = \sigma^2$.

Note. $b_T$ depends on $\theta$ and perhaps other unknowns.
Example: Uniform Distributions
Estimating the Number of German Tanks

If $X_1, \cdots, X_n \sim \text{Unif}[0, \theta]$, then

$$\hat{\theta}_{\text{MOM}} = 2\bar{X}$$

is unbiased, since $E(\bar{X}) = \mu = \theta/2$. The MLE is

$$\hat{\theta}_{\text{MLE}} = \max\{X_1, \cdots, X_n\}.$$  

is biased, but

$$\hat{\theta}_{\text{BC}} = \frac{(n + 1)\hat{\theta}_{\text{MLE}}}{n}$$

is unbiased.

Variance and MSE

**Defs:** Let $T$ be an estimator of $\theta$. Then variance of $T$ is $\sigma_T^2$, measures the variability in $\theta$. The mean squared error is

$$MSE_T = E[(T - \theta)^2].$$

**Important Relation:**

$$MSE = \sigma_T^2 + b_T^2.$$  

**Comparison:** $T_1$ is better than $T_2$ if

$$MSE_{T_1} \leq MSE_{T_2}$$

always with strict $<$ sometime.

**Note:** Again, $\sigma_T^2$ and $MSE_T$ depend on $\theta$ and perhaps other unknowns.

Efficiency

**Def:** If $T_1$ and $T_2$ are estimators of $\theta$, then

$$\text{eff}(T_1, T_2) = \frac{MSE_{T_2}}{MSE_{T_1}}$$

is called the efficiency of $T_1$ with respect to $T_2$.

**Note:** It $T_1$ and $T_2$ are unbiased estimators, then

$$\text{eff}(T_1, T_2) = \frac{\sigma_{T_2}^2}{\sigma_{T_1}^2}.$$  

**Example:** In the uniform example,

$$\lim_{n \to \infty} \text{eff}(\hat{\theta}_{\text{MLE}}, \hat{\theta}_{\text{BC}}) = \frac{1}{2}.$$  

Regular Models

**Def:** The model is regular if $\ell(\theta|x)$ is differentiable in $\theta$ and $\cdots$

**Examples:** Regular Models

- a) Binomial$(n, p)$
- b) Exponential$(\lambda)$
- c) Normal$(\mu, \sigma^2)$
- d) Poisson$(\lambda)$

**Example:** A Non-regular model: Uniform$(0, \theta)$.  

The Score Function

Consider a regular model $X \sim f(x; \theta)$ and let
\[
\ell(\theta|x) = \log[f(x; \theta)].
\]
Then
\[
\ell'(\theta|x) = \frac{d}{d\theta} \ell(\theta|x)
\]
is called the score function; and
\[
\ell'(\theta|x) = 0
\]
is called the likelihood equation. Thus, the maximum likelihood estimator (often) solves the likelihood equation. The second derivative
\[
\ell''(\theta|x) = \frac{d^2}{d\theta^2} \ell(\theta|x)
\]
measures the stability of (numerical) solutions to the LE.

Properties

For a regular model,
\[
\ell(\theta|x) = \log[f(x; \theta)],
\ell'(\theta|x) = \frac{f'(x; \theta)}{f(x; \theta)},
\ell''(\theta|x) = \frac{f''(x; \theta)}{f(x; \theta)} - \ell'(\theta|x)^2.
\]
Further,
\[
E[\ell'(\theta|X)] = 0,
E[\ell'(\theta|X)^2] = -E[\ell''(\theta|X)].
\]

Information

Def: For a regular model,
\[
i(\theta) = E[\ell'(\theta|X)^2].
\]
Then $i(\theta)$ is called the information. Alternatively,
\[
i(\theta) = -E[\ell''(\theta|X)].
\]

Theorem: The Cramer Rao Inequality. If $T$ is unbiased, then
\[
\sigma_T^2 \geq \frac{1}{i(\theta)}.
\]
Def: $T$ is said to be efficient, if there is equality (for all $\theta$).

For example, in the discrete case,
\[
E[\ell'(\theta|X)] = \sum_{x \in X} \ell'(\theta|x)f(x; \theta)
= \sum_{x \in X} f'(x; \theta)
= \frac{d}{d\theta} \sum_{x \in X} f(x; \theta)
= \frac{d}{d\theta} 1
= 0,
\]
and the other assertion may be proved similarly.
Example

Poisson($\lambda$)

If

\[ X_1, \ldots, X_n \sim \text{ind Poisson}(\lambda), \]

then

\[
L(\lambda|x) = \prod_{k=1}^{n} \frac{1}{x_k!} \lambda^{x_k} e^{-\lambda} \\
= C \lambda^{(x_1 + \cdots + x_n)} e^{-n\lambda},
\]

where

\[ C = 1/(x_1! \times \cdots \times x_n!). \]

So,

\[
\ell(\lambda|x) = (x_1 + \cdots + x_n) \log(\lambda) - n\lambda + \log(C),
\]

and

\[
\ell'(\lambda|x) = \frac{x_1 + \cdots + x_n}{\lambda} - n, \\
\ell''(\lambda|x) = - \frac{x_1 + \cdots + x_n}{\lambda^2},
\]

iff

\[
\hat{\lambda} = \frac{x_1 + \cdots + x_n}{n} = \bar{x}.
\]

For the information

\[
-E[\ell''(\lambda|X)] = \frac{1}{\lambda^2} E(X_1 + \cdots + X_n)
\]

\[
= \frac{n\lambda}{\lambda^2}
\]

\[ = \frac{n}{\lambda}. \]

In this example,

\[
E(\hat{X}) = \lambda
\]

and

\[
\sigma_{\hat{X}}^2 = \frac{\lambda}{n} = \frac{1}{n i(\lambda)}. \]

So, $\hat{X}$ is efficient.

Asymptotic Efficiency of the MLE

Guiding Principle of Statistics: For regular models and large samples (large $|\ell''(\hat{\theta})| >> 1$),

\[
\hat{\theta}_{MLE} \approx \text{Normal} \left[ \theta, \frac{1}{i(\theta)} \right]
\]

Notes 1. Cannot be applied blindly.

2. The basis for many statistical procedures.
Example: The Population Median

The Problem. Suppose
\[ X_1, \ldots, X_n \sim \text{ind Normal}[\mu, \sigma^2]. \]
Estimate the median of the \( X_i \), \( \theta \) say.

Intuitive: Suppose \( n = 2m + 1 \) and let \( X_{(1)} < \cdots < X_{(n)} \) be the order statistics. Let
\[ \tilde{\theta} = X_{(m+1)}, \]
the sample median

MLE: For an normal distribution, the population mean is the population median. So, the MLE is
\[ \hat{\theta} = \bar{X}. \]

Here
\[ \hat{\theta} \sim \text{Normal}[\theta, \frac{\sigma^2}{n}]. \]
It can be shown that
\[ \hat{\theta} \approx \text{Normal}[\theta, \frac{\pi \sigma^2}{2n}] \]
So,
\[ \text{eff}(\hat{\theta}, \tilde{\theta}) \approx \frac{\pi \sigma^2 / 2n}{\sigma^2 / n} = \frac{\pi}{2}. \]

Review: The Score Function

Consider a regular model \( X \sim f(x; \theta) \) and let
\[ \ell(\theta|x) = \log[f(x; \theta)]. \]
Then
\[ \ell'(\theta|x) = \frac{d}{d\theta} \ell(\theta|x) \]
is called the score function; and
\[ \ell'(\theta|x) = 0 \]
is called the likelihood equation. Thus, the maximum likelihood estimator (often) solves the likelihood equation. The second derivative
\[ \ell''(\theta|x) = \frac{d^2}{d\theta^2} \ell(\theta|x) \]
measures the stability of (numerical) solutions to the LE.

Review: Information

Def: For a regular model,
\[ i(\theta) = E[\ell'(\theta|X)^2]. \]
Then \( i(\theta) \) is called the information. Alternatively,
\[ i(\theta) = -E[\ell''(\theta|X)]. \]

The Cramer Rao Inequality. If \( T \) is unbiased, then
\[ \sigma_T^2 \geq \frac{1}{i(\theta)}. \]
Def: \( T \) is said to be efficient, if there is equality (for all \( \theta \)).
Asymptotic Efficiency of the MLE

Guiding Principle of Statistics: For regular models and large samples (large $|\ell''(\hat{\theta})| > 1$),

\[ \hat{\theta}_{\text{MLE}} \approx \text{Normal} \left[ \theta, \frac{1}{i(\theta)} \right] \]

Notes 1. Cannot be applied blindly.

2. The basis for many statistical procedures.

Example: Genetics

Pairs of plants were crossed, and characteristics of the leaves of the offspring recorded. The possible outcomes were

<table>
<thead>
<tr>
<th>Type</th>
<th>Probability</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starchy Green</td>
<td>$p_1$</td>
<td>$X_1$</td>
</tr>
<tr>
<td>Starchy White</td>
<td>$p_2$</td>
<td>$X_2$</td>
</tr>
<tr>
<td>Sugary Green</td>
<td>$p_3$</td>
<td>$X_3$</td>
</tr>
<tr>
<td>Sugary White</td>
<td>$p_4$</td>
<td>$X_4$</td>
</tr>
</tbody>
</table>

Note: $X_1 + X_2 + X_3 + X_4 = n$, the total number of offspring.

Model: Assuming independence of the offspring,

\[(X_1, X_2, X_3, X_4) \sim \text{Multinomial}(n; p_1, p_2, p_3, p_4);
\]

that is

\[ P[X = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4] = \frac{n!}{x_1! \times \cdots \times x_4!} p_1^{x_1} \times p_2^{x_2} \times p_3^{x_3} \times p_4^{x_4}. \]

Genetic Theory: From genetic theory,

\[ p_1 = (2 + \theta)/4, \]

\[ p_2 = p_3 = (1 - \theta)/4, \]

\[ p_4 = \theta/4 \]

for some $0 \leq \theta \leq 1$.

The Problem: Estimate $\theta$.

The Likelihood Function: Let

\[ f(x; \theta) = P[X = x] = \frac{n!}{x_1! \times \cdots \times x_4!} p_1^{x_1} \times p_2^{x_2} \times p_3^{x_3} \times p_4^{x_4}, \]

where $x = (x_1, x_2, x_3, x_4)$. Then

\[ L(\theta|x) = f(x; \theta) \]

and

\[ \ell(\theta|x) = x_1 \log(2 + \theta) + (x_2 + x_3) \log(1 - \theta) + x_4 \log(\theta) + C, \]

where $C$ depends on $x$ (and not on $\theta$).

The MLE:

\[ \ell(\hat{\theta}|x) = \max_{\theta} \ell(\theta|x) \]
Example

$x_1 = 1997, \ x_2 = 904, \ x_3 = 906, \ x_4 = 32$

The Score Function: Here

$$\ell'(\theta|x) = \frac{x_1}{2 + \theta} - \frac{x_2 + x_3}{1 - \theta} + \frac{x_4}{\theta}.$$  

and

$$\ell'(\theta|x) = 0$$

iff

$$A\theta^2 + B\theta + C = 0,$$

where

$$A = n$$
$$B = x_1 + 2(x_2 + x_3) - x_4,$$
$$C = -2x_4$$

The MLE: So,

$$\hat{\theta} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}.$$  

Example: In More Detail

$$\hat{\theta} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}.$$  

Example: In More Detail

$$\hat{\theta} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}.$$  

The MLE: So,

$$\hat{\theta} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}.$$  

The MLE: So,

$$\hat{\theta} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}.$$  

Example: In the example, $\hat{\theta} = 0.0357$.

Question: How accurate is this estimate.

From Statistical Theory:

$$\hat{\theta} \approx \text{Normal}[\theta, \frac{1}{i(\theta)}],$$

where $i(\theta)$ is the information.

Question: How informative was the experiment.
The Information

Here

\[
\ell''(\theta|x) = - \left[ \frac{x_1}{(2+\theta)^2} + \frac{x_2 + x_3 + x_4}{(1-\theta)^2} + \frac{x_4}{\theta^2} \right],
\]

\[
E(X_1) = np_1 = \frac{(2+\theta)n}{4},
\]

\[
E(X_2) = np_2 = \frac{(1-\theta)n}{4} = E(X_3),
\]

\[
E(X_4) = np_4 = \frac{\theta n}{4}
\]

So,

\[
i(\theta) = -E[\ell''(\theta|X)] = \left[ \frac{(2+\theta)n}{4(2+\theta)^2} + 2\frac{(1-\theta)n}{4(1-\theta)^2} + \frac{\theta n}{4\theta^2} \right]
\]

and

\[
i(\theta) = \frac{n(1+2\theta)}{2(2+\theta)(1-\theta)\theta}
\]

So,

\[
P[|\hat{\theta} - \theta| \leq \frac{2}{\sqrt{i(\theta)}}] \approx .954.
\]

For the given data set \(\hat{\theta} = .0357\) and \(2/\sqrt{i(\hat{\theta})} = .0114\). So, the estimate can be reported as

\[
\hat{\theta} \pm \frac{2}{\sqrt{i(\hat{\theta})}} = .0357 \pm .0114.
\]