Point Estimation

Charlie Gibbons
University of California, Berkeley

ARE 210

Fall 2015

Outline

Estimators for parametric models

Consider a *parametric model*:

$$\mathcal{F} = \{f(x;\theta) \,:\, \theta \in \Theta\}\,.$$

An *estimator* of θ, $\hat{\theta}$, is some function of the data that is used to "guess" the population parameter.

Evaluating an estimator

There are three properties that we are interested in when coming up with an estimator:

- Unbiasedness
- Consistency
- Efficiency

Mean squared error

The *mean squared error* of an estimator is the difference between the estimator and the true value:

$$\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right].$$

For $\bar{\theta}=\mathbb{E}\left[\hat{\theta}\right]$, the MSE can be rewritten as:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] &= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}+\bar{\theta}-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)^2\right]+\mathbb{E}\left[\left(\bar{\theta}-\theta\right)^2\right]+2\mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)\left(\bar{\theta}-\theta\right)\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)^2\right]+\mathbb{E}\left[\left(\bar{\theta}-\theta\right)^2\right]+2\left(\bar{\theta}-\theta\right)\mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)^2\right]+\left(\bar{\theta}-\theta\right)^2.
\end{aligned}$$

Bias-variance trade-off

We have

$$\text{MSE} = \mathbb{E}\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right] + \left(\bar{\theta} - \theta\right)^2 = \text{Var}\left(\hat{\theta}\right) + \text{bias}^2.$$

Estimator of the variance

Let Y_i be i.i.d. Normal with mean μ and variance σ^2.

Example: Let $\bar{Y}$ be an estimator for μ. Is it unbiased? Find its mean.

Consider an estimator of the variance:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Is this estimator of σ^2 unbiased?

Unbiasedness of s^2

$$
\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] \\
&= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left((Y_i - \mu) - (\bar{Y} - \mu)\right)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[(Y_i - \mu)^2\right] - 2\frac{1}{n-1}\mathbb{E}\left[(\bar{Y} - \mu)\sum_{i=1}^{n}(Y_i - \mu)\right] \\
&\quad + \frac{n}{n-1}\mathbb{E}\left[(\bar{Y} - \mu)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[(Y_i - \mu)^2\right] - \frac{n}{n-1}\mathbb{E}\left[(\bar{Y} - \mu)^2\right]
\end{aligned}
$$

Unbiasedness of s^2, continued

Plugging in the variance of $\bar{Y}$ gives

$$\begin{aligned} &= \frac{n}{n-1}\text{Var}(Y_i) - \frac{n}{n-1}\frac{\text{Var}(Y_i)}{n} \\ &= \text{Var}(Y_i). \end{aligned}$$

Estimators of the variance

Consider a more general estimator $\hat{\sigma}^2 = a\frac{1}{N-1}\sum\left(Y_i - \bar{Y}\right)^2$, where a could equal 1 or $(N-1)/N$.

For $a = 1$ and Y_i i.i.d. Normal, it can be shown that

$$\text{Var}(\hat{\sigma}^2) = \frac{2}{N-1}\sigma^4.$$

Question: What is the MSE for $\hat{\sigma}^2$?

Question: Is the MSE smaller for the unbiased or the biased estimator?

Empirical CDF

Let $X_1, \ldots, X_N$ be an i.i.d. sample.

The *empirical cumulative distribution function* (ECDF) is

$$\hat{F}(x) = \frac{1}{N} \sum_i \mathbb{I}\{X_i \leq x\}.$$

Note that $\mathbb{I}\{X_i \leq x\}$ is a binary random variable; it is 1 with probability $F(x)$.

Hence, $\mathbb{I}\{X_i \leq x\} \sim \text{Bernoulli}(F(x))$.

Question: Find the expected value and variance of $\hat{F}(x)$.

Convergence

The Glivenko-Cantelli theorem states that

$$\sup_x |\hat{F}(x) - F(x)| \xrightarrow{\text{a.s.}} 0.$$

This implies *uniform* convergence across all values of x.

The *Dvoretzky-Keifer-Wolfowitz inequality* quantifies the rate of uniform convergence:

$$\Pr\left(\sup_x |\hat{F}(x) - F(x)| > \epsilon\right) \leq 2\exp\{-2n\epsilon^2\}.$$

Covariance of CDF at two points

Find the covariance of $\hat{F}(x)$ and $\hat{F}(y)$ for $x < y$.

We know that $\hat{F}$ is unbiased, so the trick is finding $\mathbb{E}\left[\hat{F}(x)\hat{F}(y)\right]$.

$$\begin{aligned}\mathbb{E}\left[\hat{F}(x)\hat{F}(y)\right] &= \frac{1}{N^2}\mathbb{E}\left[\sum_i \mathbb{I}\{Y_i \leq x\}\sum_j \mathbb{I}\{Y_j \leq y\}\right] \\ &= \frac{1}{N^2}\mathbb{E}\left[\sum_i\sum_j \mathbb{I}\{Y_i \leq x\}\mathbb{I}\{Y_j \leq y\}\right]\end{aligned}$$

Expectation, continued

$$
\begin{aligned}
&= \frac{1}{N^2} \sum_i \sum_j \mathbb{E}\left[\mathbb{I}\{ Y_i \leq x\}\mathbb{I}\{ Y_j \leq y\}\right] \\
&= \frac{1}{N^2} \left[\sum_i \mathbb{E}[\mathbb{I}\{ Y_i \leq x\}] + \sum_i \sum_{j \neq i} \mathbb{E}[\mathbb{I}\{ Y_i \leq x\}\mathbb{I}\{ Y_j \leq y\}]\right] \\
&= \frac{1}{N^2} \left[NF(x) + \sum_i \mathbb{E}[\mathbb{I}\{ Y_i \leq x\}] \sum_{j \neq i} \mathbb{E}[\mathbb{I}\{ Y_j \leq y\}]\right] \\
&= \frac{1}{N^2} \left[NF(x) + N(N-1)F(x)F(y)\right].
\end{aligned}
$$

Covariance of CDF values

Putting it together:

$$\begin{aligned}
\text{Cov}\left(\hat{F}(x), \hat{F}(y)\right) &= \mathbb{E}\left[\hat{F}(x), \hat{F}(y)\right] - \mathbb{E}\left[\hat{F}(x)\right]\mathbb{E}\left[\hat{F}(y)\right] \\
&= \frac{1}{N}\left[F(x) + (N-1)F(x)F(y)\right] - \frac{N}{N}F(x)F(y) \\
&= \frac{1}{N}F(x)\left[1 - F(y)\right].
\end{aligned}$$

Statistical functionals

A *statistical functional* $T(F)$ is any function of the CDF F.
Example: the median

A linear functional can be written $T(F) = \int r(x)\, dF(x)$.
Example: the mean and other moments

The *plug-in principle* states that we can estimate $T(F)$ according to $T\left(\hat{F}\right)$.

Method of moments

The *method of moments* prescribes using (at least) k moments to estimate a k dimensional parameter vector θ.

Example: What are the method of moments estimators for the mean and variance of a Normal distribution?

Example: What is the method of moments estimator for the mean and variance of a Poisson random variable?

Binomial example

Let X_i have a binomial distribution

$$\Pr(X_i = x) = \binom{k}{x} p^x (1-p)^{k-x}$$

with x known, but k and p unknown.

MOM estimators

We know that:

$$\mathbb{E}[X_i] = kp$$
$$\mathbb{E}[X_i^2] = kp(1-p) + (kp)^2.$$

The method of moments estimators are:

$$\bar{X} = \hat{k}\hat{p}$$
$$\frac{1}{n}\sum_i X_i^2 = \hat{k}\hat{p}(1-\hat{p}) + (\hat{k}\hat{p})^2.$$

Solving for $\hat{k}$

This gives:

$$\frac{1}{n}\sum_i X_i^2 = \hat{k}\hat{p}(1-\hat{p}) + (\hat{k}\hat{p})^2$$

$$\frac{1}{n}\sum_i X_i^2 = \bar{X}(1-\hat{p}) + \bar{X}^2$$

$$\frac{1}{n}\sum_i X_i^2 - \bar{X}^2 = \bar{X}(1-\hat{p})$$

$$\frac{\frac{1}{n}\sum_i \left(X_i - \bar{X}\right)^2}{\bar{X}} = 1 - \frac{\bar{X}}{\hat{k}}$$

$$\frac{\bar{X} - \frac{1}{n}\sum_i \left(X_i - \bar{X}\right)^2}{\bar{X}} = \frac{\bar{X}}{\hat{k}}$$

$$\hat{k} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_i \left(X_i - \bar{X}\right)^2}.$$

Commentary

And

$$\hat{p} = \frac{\bar{X}}{\hat{k}}.$$

Note that the denominator of $\hat{k}$ need not be positive, making $\hat{k}$ and $\hat{p}$ negative, which is impossible.

The method of moments is not *range preserving*.

Maximum likelihood estimation

Suppose that X_i are i.i.d. with PDF $f(x;\theta)$.

What's the probability that we observe the sample $X_1 = x_1, \ldots X_n = x_n$?

$$\begin{aligned} f(X_1 = x_1, \ldots, X_n = x_n; \theta) &= f(X_1 = x_1; \theta) \cdots f(X_n = x_n; \theta) \\ &= \prod_i f(x_i; \theta). \end{aligned}$$

Suppose that we choose a $\hat{\theta}$ such that witnessing this particular sample is as likely as possible; this is *maximum likelihood estimation*.

Maximizing *given* θ

Note: We choose a $\hat{\theta}$ such that

$$f(X_1 = x_1, \ldots, X_n = x_n; \hat{\theta})$$

is maximized; we choose the $\hat{\theta}$ that makes the data most likely to have been observed.

This is not the same as finding the $\hat{\theta}$ that maximizes

$$f(\hat{\theta}; X_1 = x_1, \ldots, X_n = x_n);$$

$\hat{\theta}$ is not the most likely value given the data.

This latter goal is the approach of *Bayesian statistics*.

Likelihood function

The *likelihood function* is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

We choose $\hat{\theta}$ by maximizing the likelihood

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}_n(\theta).$$

Log likelihood function

Recall that, for any monotonically increasing function $g(\cdot)$,

$$\underset{z}{\operatorname{argmax}} f(z) = \underset{z}{\operatorname{argmax}} g(f(z)).$$

It is almost always easier to maximize the *log likelihood function*

$$\ell_n(\theta) = \sum_{i=1}^{n} \log\left(f(x_i;\theta)\right).$$

Define the *score* as the derivative of the log likelihood:

$$s(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log\left(f(x_i;\theta)\right)}{\partial \theta} = \sum_{i=1}^{n} s_i(\theta).$$

Typical steps for finding the MLE

The usual method for finding the MLE is to:

1. Find the log likelihood function.
2. Take the (partial) derivative(s) of the log likelihood function with respect to θ to find the score function.
3. Set the score equal to 0 and solve for $\hat{\theta}$.
4. Ensure that we've found the maximum by checking that
   - the second derivative for a univariate θ is negative or
   - at least one second derivative is negative and the determinant of the Hessian is positive for the two variable case.
5. Check the boundary conditions to ensure that a maximum is not found there.

Regularity conditions

To derive the properties of the MLE, we will assume the following regularity conditions:

- The PDF of x_i is *identified*: $f(x_i; \theta_1) \neq f(x_i; \theta_2) \, \forall \, \theta_1 \neq \theta_2$.
- $\theta \in \Theta$, where Θ is finite dimensional, closed, and compact.
- The (log) likelihood has at least three continuous and bounded derivatives (typically two will be sufficient).
- The order of integration and differentiation of the (log) likelihood can be reversed; in particular, this requires that the limits of integration do not depend upon θ.

Notation

We will derive the properties of the MLE under possible *misspecification* of the likelihood (though still assuming that the X_i are i.i.d.).

To derive the properties of the maximum likelihood estimator:

- Make a modeling assumption that $X_i \sim F_i$.
- The true DGP is $X_i \sim F_i^*$.
- Find $\hat{\theta} : \sum_{i=1}^{n} s_i\left(\hat{\theta}\right) = 0$.
- Define $\theta_* : \mathbb{E}_* \left[\sum_{i=1}^{n} s_i(\theta) \right] = 0$ (the *pseudo-true* value of θ).

Taylor expansion

Applying a Taylor expansion to our score condition gives

$$0 = \sum_{i=1}^{n} s_i\left(\hat{\theta}\right) \approx \sum_{i=1}^{n} s_i(\theta_*) + \sum_{i=1}^{n} \left.\frac{\partial s_i(\theta)}{\partial \theta'}\right|_{\theta=\theta_*} \left(\hat{\theta} - \theta_*\right).$$

Rearranging and multiplying by $\sqrt{n}$ gives

$$\sqrt{n}\left(\hat{\theta} - \theta_*\right) = \left[-\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial s_i(\theta)}{\partial \theta'}\right|_{\theta=\theta_*}\right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i(\theta_*).$$

Fisher information

Let

$$A = -\operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial s_i(\theta)}{\partial \theta'} \right|_{\theta=\theta_*} = -\mathbb{E}_* \left[\frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial s_i(\theta)}{\partial \theta'} \right|_{\theta=\theta_*} \right].$$

A is called the *Fisher information matrix.*

The Hessian

Recall that $s_i(\theta) = \frac{\partial \log(f(x_i;\theta))}{\partial \theta}$, which means that

$$\frac{\partial s_i(\theta)}{\partial \theta'} = \frac{\partial^2 \log\left(f(x_i;\theta)\right)}{\partial \theta \partial \theta'}.$$

The Fisher information matrix is the negative expected value of the *Hessian*, the matrix of second-order partial derivatives of, in our case, the log likelihood function:

$$H(\theta) = \begin{bmatrix} \frac{\partial \ell_n(\theta)}{\partial \theta_1^2} & \frac{\partial \ell_n(\theta)}{\partial \theta_1 \theta_2} & \cdots & \frac{\partial \ell_n(\theta)}{\partial \theta_1 \theta_n} \\ \frac{\partial \ell_n(\theta)}{\partial \theta_2 \theta_1} & \frac{\partial \ell_n(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial \ell_n(\theta)}{\partial \theta_2 \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ell_n(\theta)}{\partial \theta_n \theta_1} & \cdots & \cdots & \frac{\partial \ell_n(\theta)}{\partial \theta_n^2} \end{bmatrix}.$$

Further convergence

Additionally, let:

$$B = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i(\theta) s_j(\theta)' \big|_{\theta=\theta_*}$$

$$= \mathbb{E}_* \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} s_i(\theta) s_j(\theta)' \big|_{\theta=\theta_*} \right]$$

Under the regularity conditions and correct specification (*i.e.*, $F_i = F_i^*$), then $A = B$. This is the *information matrix equality.*

Applying the CLT

The CLT tells us that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i \left(\hat{\theta} \right) \xrightarrow{d} \mathrm{N}(0, B).$$

Then,

$$\sqrt{n} \left(\hat{\theta} - \theta_* \right) \xrightarrow{d} \mathrm{N}(0, A^{-1} B A^{-1}).$$

Under the regularity conditions and correct specification, this simplifies to

$$\sqrt{n} \left(\hat{\theta} - \theta_* \right) \xrightarrow{d} \mathrm{N}(0, A^{-1}).$$

MLE for normal random variables

Example: For a Normal distribution, find:

- The log likelihood,
- The score, and
- The Hessian.

Propose MLE estimators for μ and σ^2. Show that this is a maximum.

MLE of functions of θ

Suppose that we want to estimate $g(\theta)$ and let G be the image of Ω under the function g. For each $t \in G$, define $G_t = \{\theta : g(\theta) = t\}$ and

$$\ell^*(t) = \max_{\theta \in G_t} \ell_n(\theta).$$

Let the MLE of $g(\theta)$ be

$$\hat{t} = \operatorname*{argmax}_{t \in G} \ell^*(t).$$

Invariance

Since $\ell^*(t)$ is a maximum over a subset of Ω while $\hat{\theta}$ is the maximum over all of Ω, $\ell^*(t) \leq \ell\left(\hat{\theta}\right)$ for all $t \in G$.

Let $\hat{t} = g\left(\hat{\theta}\right)$. See that $\hat{\theta} \in G_{\hat{t}}$.

Since $\hat{\theta}$ maximizes ℓ over all θ, it maximizes ℓ over $\theta \in G_{\hat{t}}$. Hence, $\ell^*\left(\hat{t}\right) = \ell\left(\hat{\theta}\right)$ and $\hat{t} = g\left(\hat{\theta}\right)$ is the MLE of $g(\theta)$.

This is the *invariance* property of MLE.

Binomial example revisited

Recall the MOM example of $X_i \sim \text{Binomial}(k, p)$, with k and p unknown.

We noted that the MOM was not range preserving; the MLE is.

Suppose that $X = (16, 18, 22, 25, 27)$. The MLE of k is 99.

Suppose that $X = (16, 18, 22, 25, 28)$. The MLE of k is 190.

The MLE can be very sensitive to the underlying data.

Non-uniqueness

Suppose that $X_i \sim \text{Uniform}[\theta, \theta + 1]$. Find the MLE for θ.

We can write the likelihood as

$$f_n(x; \theta) = \begin{cases} 1 & \text{if } \theta \leq x_i \leq \theta + 1 \quad \forall x_i \\ 0 & \text{otherwise.} \end{cases}$$

Alternatively,

$$f_n(x; \theta) = \begin{cases} 1 & \max\{x_i\} - 1 \leq \theta \leq \min\{x_i\} \\ 0 & \text{otherwise.} \end{cases}$$

Since the likelihood is the same for any value of θ satisfying $\max\{x_i\} - 1 \leq \theta \leq \min\{x_i\}$, any of these could be the MLE; in this case, the MLE is not unique.

Mixture distribution

Suppose that $X_i \sim \mathrm{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ with probability $1/2$ and $X_i \sim \mathrm{N}(0, 1)$ with probability $1/2$. Then

$$f(x; \mu, \sigma^2) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) + \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)\right].$$

Consider the case where $\mu = x_1$ and $\sigma^2 \to 0$. Then, for $i \neq 1$,

$$f(x \mid \mu, \sigma^2) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) + \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - x_1)^2}{2\sigma^2}\right)\right].$$

For $f(x_1 \mid \mu, \sigma^2) \to \infty$.

Non-existence

We can make the likelihood be arbitrarily large by pushing σ^2 to 0. But σ^2 must be larger than 0. Here, the MLE does not exist.

If we let $\sigma^2 = 0$, then we have n MLEs: the pairs of $\mu = \{x_i\}$ and $\sigma^2 = 0$.

Integral tricks

We know that

$$\int f(x;\theta)\,dx = 1.$$

Differentiate both sides with respect to θ:

$$\nabla_\theta \int f(x;\theta)\,dx = 0.$$

Assume that we can change the order of integration and differentiation to give

$$\int \nabla_\theta f(x;\theta)\,dx = 0.$$

Relating MLE and MOM

Multiply and divide by $f(x;\theta)$ to give

$$\int \frac{\nabla_\theta f(x;\theta)}{f(x;\theta)} f(x;\theta)\, dx = \int \nabla_\theta \log\left(f(x;\theta)\right) f(x;\theta)\, dx = \mathbb{E}\left[\nabla_\theta \log\left(f(x;\theta)\right)\right]$$

We can use the method of moments to estimate

$$\mathbb{E}\left[\nabla_\theta \log\left(f(x;\theta)\right)\right] \approx \frac{1}{N}\sum_i \nabla_\theta \log\left(f\left(x_i;\hat{\theta}\right)\right) = 0.$$

But this is the MLE!