10/10: Estimation

Estimation

Up to now we've been talking about random variables, distributions, etc. But where to the distributions come from? The estimation or "learning" problem is to select one or more distributions from a set of candidates on the basis of available data — and extrinsic assumptions about the problem.

Inference vs. Learning

Estimation/Learning: Selecting parameters, a distribution over parameters, or a set of cdf's for a statistical problem based on data.

Inference: Making predictions, computing statistics, expectations, or marginal probabilities for a statistical model that has already been estimated/learned.

Parameters

A statistical family with a finite collection of adjustable parameters is the starting point for a *parametric* estimation problem.

If there are an infinite number of adjustable parameters– typically entire functions or cdf's, then the problem (or approach) is said to be *non-parametric.*

Parametric vs. Non-Parametric

This can be confusing, since often "non-parametric" problems seem to have many more "parameters" than a typical parametric problem.

Non-parametric approaches make fewer assumptions about the form or "shape" of the distribution being estimated.

However, the distinction is sometimes subtle (e.g., neural nets)

Basic Concepts

Point estimation: choose a *single* parameter $\hat{\theta}$ or cdf, or other prediction.

Note: $\hat{\theta}$ is a random variable, since it is a function of the data (which is random):

$$\hat{\theta}_n = g(X_1, X_2, \ldots, X_n)$$

where g *represents an algorithm* for computing the point estimate.

Bias

The *bias* of a point estimator is

$$\text{bias}(\hat{\theta}_n) \quad = \quad E_F[\hat{\theta}_n] - \theta$$

An estimator is *unbiased* if

$$E_F[\hat{\theta}_n] = \theta$$

where $X_1, X_2, \ldots, X_n$ are iid $\sim F$.

Consistency

A point estimate of a parameter θ is *consistent* if

$$\hat{\theta}_n \longrightarrow \theta$$

(in probability). The *standard error* is the standard deviation of $\hat{\theta}_n$:

$$\text{se}(\hat{\theta}_n) \quad = \quad \sqrt{E_F(\hat{\theta}_n - E_F(\hat{\theta}_n))^2}$$

For an unbiased estimator this is

$$\text{se}(\hat{\theta}_n) \quad = \quad \sqrt{E_F(\hat{\theta}_n - \theta)^2}$$

Note that since the expectation is the "true" expectation over the data, this is in general impossible to compute.

Example

Let $X_i \sim \text{Bernoulli}(p)$.

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

satisfies

$$E[\hat{\theta}_n] = \frac{1}{n} \cdot np = p$$

so this is an unbiased estimate of $p = \theta$.

Example (cont.)

The standard error is

$$\begin{aligned} \text{se}(\hat{\theta}_n) &= \sqrt{E\left(\left(\frac{1}{n}\sum X_i\right)^2 - \theta^2\right)} \\ &= \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{\theta(1-\theta)}{n}} \end{aligned}$$

and so can't be computed. The estimated standard error is

$$\hat{\text{se}} = \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$$

Example (cont.)

Note that for the Bernoulli the estimated standard error is also given by

$$\begin{aligned} \hat{\text{se}} &= \sqrt{\frac{1}{n^2}\sum_i (X_i - \hat{\theta}_n)^2} \\ &= \sqrt{\frac{1}{n}\left(\frac{1}{n}\sum_i X_i^2 - \hat{\theta}_n^2\right)} \\ &= \sqrt{\frac{1}{n}\hat{\theta}_n(1-\hat{\theta}_n)} \end{aligned}$$

MSE

The *mean squared error* (MSE) of an estimator is defined as

$$E[(\hat{\theta}_n - \theta)^2]$$

Another way of looking at this is

$$\begin{aligned} MSE &= E[(\hat{\theta}_n - \theta)^2] \\ &= E[((\hat{\theta}_n - E[\hat{\theta}_n]) + (E[\hat{\theta}_n] - \theta))^2] \\ &= \text{Var}(\hat{\theta}_n) + \text{bias}^2(\hat{\theta}_n) \end{aligned}$$

Fundamental tradeoff.

Asymptotically Normal

An estimator is *asymptotically normal* in case

$$\frac{\hat{\theta}_n - \theta}{\mathsf{se}(\hat{\theta}_n)} \rightsquigarrow \mathcal{N}(0,1)$$

Point Estimation for Parametric Families

We have a family $\mathcal{F} = \{f_\theta(x), \theta \in \Theta\}$ and want to estimate certain parameters of interest.

Method of Moments

Suppose that $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. For $1 \leq j \leq k$, the *j-th moment* is

$$\alpha_j(\theta) \;=\; E_\theta(X^j)$$

and the j-th sample moment is

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Method of Moments

The method of moments estimator is given by the value $\hat{\theta}_n$ that solves the following system of k equations in k unknowns:

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ &\cdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k \end{aligned}$$

Method of Moments (cont.)

For the Bernoulli, MoM gives

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Method of Moments (cont.)

For the Gaussian, MoM gives

$$\begin{aligned}
\alpha_1(\hat{\theta}_n) &= \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \\
\alpha_2(\hat{\theta}_n) &= \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2
\end{aligned}$$

Leading to

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_i X_i \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_i (X_i - \hat{\mu})^2
\end{aligned}$$

Maximum Likelihood

The most commonly used method for point estimation. Given a family $\mathcal{F} = \{f(x \,|\, \theta)\}$ and data $X_1, X_2, \ldots, X_n$, the *likelihood function* is defined as

$$\mathcal{L}_n(\theta) = \prod_i f(X_i \,|\, \theta)$$

and the *log-likelihood function* is given by

$$\begin{aligned}
\ell_n(\theta) &= \log \mathcal{L}_n(\theta) \\
&= \sum_i \log f(X_i \,|\, \theta)
\end{aligned}$$

Maximum Likelihood

The *maximum likelihood estimator* is

$$\hat{\theta} = \arg\max_{\Theta} \ell_n(\theta)$$

(whenever this exists)

Maximum Likelihood

Examples (on the board): binomial, multinomial, normal, and uniform.

Note: This presentation follows Larry Wasserman's manuscript. DeGroot is heavy on the the Bayesian perspective in the early sections of chapter 6.