METHODS OF ESTIMATION

1. Introduction

Suppose we have an unknown parameter θ and have observed some data $X_1, ..., X_n$ assumed to be i.i.d. with a distribution depending on θ, and suppose we want to estimate some function $g(\theta)$. Often, simply $g(\theta) = \theta$. If the distribution is entirely determined by θ it will be written P_θ. Let $T = T(X_1, \ldots, X_n)$ be a statistic that may be used to estimate $g(\theta)$. There are several criteria or methods for choosing estimators.

2. Mean-squared error

Here are two simple facts on minimizing mean-squared errors.

Proposition 1. *(a) For any random variable X with $E(X^2) < +\infty$, the unique constant x which minimizes $E((X-c)^2)$ is $c = EX$.*
(b) For any real values $X_1, ..., X_n$, the sum $\sum_{j=1}^n (X_j - t)^2$ is minimized with respect to t for $t = \overline{X}$.

Proof. (a) We have $E((X-c)^2) = E(X^2) - 2cEX + c^2$, which goes to $+\infty$ when $c \to \pm\infty$, so to find a minimum one can find where the derivative with respect to c is 0, $-2EX + 2c = 0$, so $c = EX$.
(b) Similarly, $\sum_{j=1}^n (X_j - t)^2 = \sum_{j=1}^n X_j^2 - 2t\sum_{j=1}^n X_j + nt^2$ which goes to $+\infty$ when $t \to \pm\infty$. The derivative gives $2nt - 2\sum_{j=1}^n X_j = 0$ so $t = (\sum_{j=1}^n X_j)/n = \overline{X}$ as stated. □

When there are P_θ depending only on θ, let E_θ be expectation when θ is the true value of the parameter. The mean-squared error (MSE) of T as an estimator of $g(\theta)$, at θ, is defined as $E_\theta((T(X) - g(\theta))^2)$. One would like to make MSE's as small as possible, but in general, there is no way to choose $T(X)$ to minimize $E_\theta((T(X) - g(\theta))^2)$ for all θ. To see that, let c be any value such that $g(\theta_0) = c$ for some θ_0. Then the trivial estimator $T \equiv c$ minimizes the MSE for $\theta = \theta_0$, while for other values of θ, the estimator can have large MSE.

Define the *bias* $b(\theta) := b_T(\theta)$ of T as an estimator of $g(\theta)$ to be $b_T(\theta) := E_\theta T - g(\theta)$. For a given value of θ, a statistic T has a variance defined by $\mathrm{Var}_\theta(T) = E_\theta((T - E_\theta T)^2)$. We then have for any statistic

Date: 18.650, Sept. 25, 2014.

T such that $E_\theta(T^2) < +\infty$ for all θ, and function $g(\theta)$, that the MSE equals the variance plus the bias squared:

$$E_\theta((T - g(\theta))^2) = \mathrm{Var}_\theta(T) + b_T(\theta)^2, \tag{1}$$

because in $E_\theta([(T(X) - E_\theta T) + (E_\theta T - g(\theta))]^2)$ we have $E_\theta((T(X) - E_\theta T)(E_\theta T - g(\theta)) = 0$, as for fixed θ, the latter factor is a constant.

Equation (1) is sometimes called the "bias-variance tradeoff". In minimizing the MSE one would like both the variance and the bias to be small. In an older tradition, one first looked for estimators for which the bias is 0, then tried to minimize their variance. That does not always work well, however, as we'll see.

2.1. **Unbiased estimation.** An estimator T of $g(\theta)$ is said to be *unbiased* if for all θ, $E_\theta T = g(\theta)$. In other words, the bias $b_T(\theta) = 0$ for all θ. The sample mean $\overline{X}$ is an unbiased estimator of the true mean μ for any distribution having a finite mean. For the variance we have, recalling the sample variance defined as, for $n \geq 2$,

$$s_X^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

Proposition 2. *For any $n \geq 2$ and any $X_1, ..., X_n$ i.i.d. with $E(X_1^2) < +\infty$ and so having a finite variance σ^2, $E(s_X^2) = \sigma^2$, so s_X^2 is an unbiased estimator of σ^2.*

Proof. Let $\mu = EX_1$ and let $Y_j := X_j - \mu$ for $j = 1, ..., n$. Then Y_j are i.i.d. with the same variance σ^2. We have $\overline{Y} = \overline{X} - \mu$ and $Y_j - \overline{Y} = X_j - \overline{X}$ for each j. Thus $s_Y^2 = s_X^2$, so we can assume that $\mu = 0$. We then have

$$\sum_{j=1}^{n}(X_j - \overline{X})^2 = \sum_{j=1}^{n} X_j^2 - 2\sum_{j=1}^{n} X_j\overline{X} + n\overline{X}^2.$$

Since $\sum_{j=1}^{n} X_j = n\overline{X}$ and $E(\overline{X}^2) = \sigma^2/n$, the expectation of the displayed sum is $n\sigma^2 - n(\sigma^2/n) = (n-1)\sigma^2$. The statement follows. □

3. Maximum likelihood estimation

Let $f(x, \theta)$ be a family of probability densities or mass functions indexed by a parameter θ. Given $X_1, ..., X_n$ assumed to be i.i.d. $f(x, \theta)$,

we can form the *likelihood function*

$$f(X,\theta) := \prod_{j=1}^{n} f(X_j,\theta). \tag{2}$$

A *maximum likelihood estimator* of θ, depending on X, is a value of θ that maximizes $f(X,\theta)$, called *the* maximum likelihood estimator (MLE) if it is unique, and then a function $T(X)$ of X.

3.1. **The binomial case.** An easy case of maximum likelihood estimation is for the binomial parameter p with $0 \le p \le 1$. Suppose we observe X successes in n independent trials with probability p of success on each. The likelihood function is $\binom{n}{X}p^X(1-p)^{n-X}$. The binomial coefficient $\binom{n}{X}$ doesn't affect the maximization so let's omit it. If $X = 0$ we get $(1-p)^n$, maximized when $p = 0$. If $X = n$ it is p^n, maximized when $p = 1$. For $0 < X < n$, the likelihood function approaches 0 when $p \to 0$ or 1, so to find a maximum in $0 < p < 1$, setting the derivative with respect to p equal to 0 gives

$$\begin{aligned} 0 &= Xp^{X-1}(1-p)^{n-X} - p^X(n-X)(1-p)^{n-X-1} \\ &= X(1-p) - (n-X)p = X - np, \end{aligned}$$

so $p = \hat{p} = X/n$, the usual and natural estimate of p, also when $X = 0$ or n. By the way $\hat{p}$ is also an unbiased estimate of p.

3.2. **The case of normal distributions.**

Proposition 3. *For $X_1, ..., X_n$ i.i.d. $N(\mu, \sigma^2)$, with $-\infty < \mu < +\infty$ and $0 < \sigma < +\infty$, the MLE of μ is $\overline{X}$, and if $n \ge 2$ the MLE of σ^2 is $\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2$.*

Proof. The likelihood function is

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{\sum_{j=1}^{n}(X_j-\mu)^2}{2\sigma^2}\right) \tag{3}$$

for $0 < \sigma < \infty$ and $-\infty < \mu < \infty$. For fixed $\sigma > 0$ and X_j, to maximize with respect to μ is equivalent to minimizing $\sum_{j=1}^{n}(X_j-\mu)^2$ which gives $\mu = \overline{X}$ by Proposition 1(b).

To maximize a likelihood, which is positive, is equivalent to maximizing its log, called the log likelihood. With probability 1, since $n \ge 2$, the X_j are not all equal, so $\sum_{j=1}^{n}(X_j - \overline{X})^2 > 0$. In this case the log likelihood, maximized with respect to μ, equals, up to a constant $(-n/2)\log(2\pi)$ which doesn't affect the maximization,

$$-n\log(\sigma) - \sum_{j=1}^{n}(X_j - \overline{X})^2/(2\sigma^2).$$

As σ decreases down to 0, the first term goes to $+\infty$, but the second term goes to $-\infty$ faster, so the log likelihood goes to $-\infty$. As σ increases up to $+\infty$, the second term goes to 0 and the first to $-\infty$. So for a maximum, set $d/(d\sigma) = 0$ and get

$$-\frac{n}{\sigma} + \sum_{j=1}^{n} (X_j - \overline{X})^2/\sigma^3.$$

Multiplying by σ^3 and solving gives

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X})^2,$$

which is therefore the MLE of σ^2 as stated. □

Proposition 3 is Example B in Rice, p. 269. MLEs of parameters of other families such as Poisson and geometric are also easy to find, where the parameter space for the geometric case is $0 < p \leq 1$, and for the Poisson case it's $0 \leq \lambda < +\infty$. In these cases, all with one-dimensional parameters, the MLE may be on the boundary of the parameter space. When it's in the interior it can be found by setting a derivative of the likelihood function, or its log, equal to 0.

4. Method of moments estimation

If a family of distributions has just a one-dimensional parameter θ, and $E_\theta X$ is a function $g(\theta)$, then the method of moments estimate of θ is to choose it, if possible, such that $\overline{X} = g(\theta)$. Applying this to a binomial (n, p) distribution, one can consider $S_n = \sum_{j=1}^{n} X_j$ where X_j are i.i.d. Bernoulli(p), i.e. $X_1 = 1$ with probability p and 0 otherwise. Then S_n is a binomial (n, p) variable. For a fixed n we called this binomial random variable X. Then $\hat{p} = X/n = S_n/n$ is the usual estimate of p. It is seen above to be an unbiased estimate and the maximum likelihood estimate. We now see that it is also the method of moments estimate.

If θ is a 2-dimensional parameter, as for normal, gamma, and beta distributions, and the mean is a function $\mu(\theta)$, while the variance is a function $\sigma^2(\theta)$, the method of moments estimate of θ is a value, if it exists and is unique, such that $\mu(\theta) = \overline{X}$ and $\sigma^2(\theta) = \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X})^2$. The latter would be the variance of a discrete distribution, which is the sum of point masses $1/n$ at each X_j, called the *empirical distribution* P_n. That may be a reason for choosing the factor $1/n$ in the method of moments.

5. Estimation of the normal variance

Given $X_1, ..., X_n$ i.i.d., assumed to be $N(\mu, \sigma^2)$ for some unknown μ and σ, $\overline{X}$ as an estimator of μ is the MLE, is unbiased, and is the method of moments estimator. For σ^2, consider estimators $c_n \sum_{j=1}^n (X_j - \overline{X})^2$. Then $c_n = 1/(n-1)$ gives an unbiased estimator of σ^2, not only for normals but for any distribution having finite variance. The MLE is given by $c_n = 1/n$ by Proposition 3 and so is the method of moments estimate.

The next fact is not at all important in itself. It illustrates further that different factors c_n may be multiplied by $\sum_{j=1}^n (X_j - \overline{X})^2$ to estimate σ^2 by different criteria: $1/(n-1)$ for unbiasedness, $1/n$ for normal MLE or method of moments, and $1/(n+1)$ in the next fact. All these c_n satisfy $nc_n \to 1$ as $n \to \infty$. Yet another choice of c_n comes up in the next section.

Proposition 4. *To minimize $E_\theta((T(X) - \sigma^2)^2)$ for $n \geq 2$, for estimators of the form $T(X) = c_n \sum_{j=1}^n (X_j - \overline{X})^2$, for any $\theta = (\mu, \sigma)$, the best value of c_n is $c_n = 1/(n+1)$.*

Proof. If Z is a $N(0,1)$ variable, to find $E(Z^4)$, one can use integration by parts. Let $\phi(z)$ be the standard normal density, $\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2)$. Then

$$E(Z^4) = \int_{-\infty}^{\infty} z^4 \phi(z) dz = -\int_{-\infty}^{\infty} z^3 d\phi(z) = 0 + 3\int_{-\infty}^{\infty} z^2 \phi(z) dz = 3.$$

It follows that $\mathrm{Var}(Z^2) = \mathrm{Var}(\chi^2(1)) = 3 - 1 = 2$, and so $\mathrm{Var}(\chi^2(d)) = 2d$ for any positive integer d.

If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ for some unknown μ and σ^2, $\sum_{j=1}^n (X_j - \overline{X})^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom, which has mean $n-1$ and variance $2(n-1)$. So the MSE of our estimator is

$$\sigma^4 E\left[\left(c_n \chi^2(n-1) - 1\right)^2\right] = \sigma^4 \left[c_n^2 \left((n-1)^2 + 2n - 2\right) - 2c_n(n-1) + 1\right]$$

$$= \sigma^4 \left[(n^2-1)c_n^2 - (2n-2)c_n + 1\right].$$

The quantity in square brackets goes to $+\infty$ as $c_n \to \pm\infty$, because $n \geq 2$, so it is minimized when its derivative is 0, $2c_n(n^2-1) - (2n-2) = 0$. Factoring out $2n - 2 > 0$ gives $c_n = 1/(n+1)$ as claimed. □

So by four different criteria, the selected values of c_n are $1/(n-1)$, $1/n$, and $1/(n+1)$ (only two of the criteria agree).

6. Inadmissibility and the variance

An estimator $T(X)$ is called *inadmissible* as an estimator of $g(\theta)$, for mean-squared error, if there is another estimator $U(X)$ such that:
(i) $E_\theta[(U(X) - g(\theta))^2] \leq E_\theta[(T(X) - g(\theta))^2]$ for all θ, and
(ii) $E_\theta[(U(X) - g(\theta))^2] < E_\theta[(T(X) - g(\theta))^2]$ for some θ.

If there is no such U then T is called *admissible*.

Surprisingly, the usual sample variance s_X^2 turned out to be inadmissible as an estimator of the true variance σ^2 under very general conditions, as Yatracos (2005) showed. Again consider estimators

$$c_n \sum_{j=1}^{n} (X_j - \overline{X})^2$$

of σ^2, where we know that $c_n = 1/(n-1)$ gives an unbiased estimator of σ^2 whenever it is finite, whereas $c_n = 1/n$ gives the maximum likelihood estimator for normal distributions and the statistic used in method-of-moments estimation. Yatracos proved the following fact: let $X_1, ..., X_n$ be i.i.d. with any distribution such that $E(X_1^4) < \infty$, X_j are not constant, and in a family such that for any c with $0 < c < \infty$, the distribution of cX_1 is also in the family. Then the classical sample variance s_X^2 with $c_n = 1/(n-1)$ is inadmissible as an estimator of the true variance. An estimator with smaller mean-squared error is obtained by taking

$$c_n = \frac{n+2}{n(n+1)}. \tag{4}$$

Of course, the resulting estimator has a non-zero bias, but the bias becomes very small as n becomes large and the reduction in variance is enough to make the total MSE smaller. In detail, Yatracos's theorem is as follows:

Theorem 1 (Yatracos)**.** *There is a constant d_n depending on n, namely $d_n = \frac{(n+2)(n-1)}{n(n+1)}$, such that for any $n \geq 2$ and for all $X_1, ..., X_n$ i.i.d. with $E(X_1^4) < +\infty$ and variance σ^2 with $\sigma > 0$, the mean-square error of $d_n s_X^2$ as an estimator of σ^2 is less than that of s_X^2, that is,*

$$E\left((d_n s_X^2 - \sigma^2)^2\right) < E(\left(s_X^2 - \sigma^2)^2\right).$$

A proof is given in www-math.mit.edu/∼rmd/466/yatracos.pdf.

It is not claimed that the factor given by (4) is in any way optimal. In fact for normal distributions we know by Proposition 4 that $1/(n+1)$ is optimal. The Yatracos estimator may itself be inadmissible. All that Yatracos's theorem says is that his factor is always better, under the given very general conditions, for purposes of estimating σ^2 with smaller mean-squared error, than the classical factor $1/(n-1)$.

REFERENCE

Yatracos, Y. (2005). Artificially augmented samples, shrinkage, and mean squared error reduction. *J. Amer. Statist. Assoc.* **100**, 1168–1175.