More on Estimation.

In the previous chapter we looked at the properties of estimators and the criteria we could use to choose between types of estimators. Here we examine more closely some very popular basic estimation techniques, two of which focus on the estimation of parameters of a pre-specified probability density function (Maximum Likelihood and Method of Moments techniques) and the third which focuses on the estimation of the shape of an un-specified probability density function (kernel estimation). In all cases we are confronted with a random sample X_i, i = 1, 2,..,n, and in the first two cases we know the form of the p.d.f. $f(x,\theta)$ but not the value of the parameter θ (often there will be more than one parameter, the techniques are readily extended to deal with this situation) in the third case we do not know the form of f all we are trying to calculate is the value of f( ) for a given x. The third case relates solely to continuous random variables, the first two cases relate to both discrete and continuous random variables, in our discussion we refer only to the continuous case though we will give examples of discrete random variable problems.

Maximum Likelihood Estimation.

The intuition behind this technique is to choose a value for the unknown θ that will make the chance of us having obtained the sample we did obtain as big as possible. The rationale for this is that any sample we get is going to be a more likely to be a high probability sample than a low probability sample. Imagine we wish to estimate the average height of males and we randomly sample 4 males from off the street, we would be surprised if all 4 were above 7 feet and similarly we would be surprised if they were all below 4 feet. This is because they are unlikely samples. We would be a lot less surprised if their heights were between 5 and 6 feet because that would constitute a more likely sample. Thus it makes sense to choose a value for θ which maximizes the probability of having got the sample that we got.

Given $f(x,\theta)$ and independently drawn X_i;s, the joint density of the sample which is referred to as L, the likelihood, is given by:

$$L = \prod_{i=1}^{n} f(X_i, \theta)$$

and the estimation technique simply amounts to deriving the formula for θ in terms of the X_i;s which maximizes this function with respect to θ. For technical reasons (i.e. the algebra is usually easier!) we usually maximize the log of the likelihood. When there is more than one parameter the first order conditions are simply solved simultaneously (see the examples below).

Method of Moments Estimation.

The motivation here is quite different from, and somewhat more straightforward than, that for the maximum likelihood method, it relies on common sense. We have seen in an earlier chapter that given f(x,θ) we can obtain a formula for the theoretical mean or expected value of x and we can similarly obtain a formula for the theoretical variance and any other moments of x, for example if x is a continuous random variable we have:

$$E(X) = \int_{-\infty}^{\infty} xf(x,\theta)dx :$$

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x,\theta)dx$$

The sample mean and sample variance are estimates of E(X) and V(X) respectively so all that is needed is to set one of the formulae to its corresponding sample equivalent and then solve for the value of θ, when there is more than one parameter we choose as many moments as we have parameters to solve for.

Some Examples.

1. The Poisson Distribution provides us with an example of a discrete distribution with one parameter. In this case X_i is an integer ≥ 0 and $f(x,\lambda)$ is of the form:

$$f(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

and $E(x) = \lambda$. In this case the Method of Moments technique is very straight forward, since we simply set our estimator of λ to the sample mean $\bar{x}$. For finding the formula for the Maximum Likelihood estimator the logarithm of the likelihood is given by:

$$\ln L = \ln \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} = \sum_{i=1}^{n} (X_i \ln \lambda - \ln(X_i!)) - n\lambda$$

and taking the derivative w.r.t. λ and setting it to 0 yields:

$$n = \sum_{i=1}^{n} \frac{X_i}{\lambda}$$

solving this for λ yields the sample mean of X, (note that in this case the Maximum Likelihood Estimator and the Method of Moments Estimator are the same).

2. The Power Function Distribution provides us with an example of a one parameter continuous distribution. In this case X_i is a number between 0 and 1 and $f(x,\theta)$ is of the form:

$$f(x,\theta) = \theta x^{\theta-1}$$

and $E(x) = \theta/(\theta+1)$. For the Method of Moments estimator we simply set the formula for E(x) equal to $\bar{x}$ and solve for θ so that our estimator for θ will be $\bar{x}/(1-\bar{x})$. For the Maximum Likelihood estimator the logarithm of the likelihood is given by:

$$\ln L = \ln \prod_{i=1}^{n} \theta X_i^{\theta-1} = \sum_{i=1}^{n} (\ln \theta + (\theta - 1) \ln X_i)$$

taking the derivative w.r.t. θ and setting to zero yields:

$$\frac{n}{\theta} = \sum_{i=1}^{n} -\ln X_i$$

re-arranging in terms of θ yields a Maximum Likelihood estimator $n/\sum -\ln(X_i)$ which of course is very different from the Method of Moments estimator above.

3. The Normal Distribution provides us with a continuous random variable example of a two parameter problem where the unknown parameters in the distribution are μ and σ^2, the mean and variance respectively. The p.d.f. in this case is given by:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $E(x) = \mu$ and $V(x) = \sigma^2$. Again the Method of Moments estimators are trivial, we simply set the estimators of μ and σ^2 equal to the sample mean and sample variance respectively. For the Maximum Likelihood estimators the logarithm of the likelihood is given by:

$$\ln L = \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{(X_i-\mu)^2}{\sigma^2}\right)^2} = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}$$

now we have to take the derivatives with respect to both μ and σ^2, set them both to zero and solve the equations simultaneously. Taking the derivatives and setting them to zero after some cancellations yields:

$$n\mu = \sum_{i=1}^{n} X_i$$

$$\frac{n}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^4}$$

solving these simultaneously yields the sample mean as the Maximum Likelihood estimator for μ (the same as the Method of Moments estimator) and $\Sigma(Xi - \bar{x})^2/n$ as the Maximum Likelihood estimator of σ^2 (which is differnt from the method of moments estimator).

Kernel Estimation.

The issue to be addressed here is the estimation of some unknown density function f(x) which underlies a sample of observations on x. Given such a sample x_i , i = 1,..,n, the generation of a naive estimate of f(x) is straightforward. If the random variable X has the density f(x) then:

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x-h < X < x+h)$$

for given h we can estimate P(x-h < X < x+h) by the proportion of observations falling into the interval x-h, x+h. Letting I( ) be an indicator function where I(z) = 1 if z is true and 0 otherwise, our estimator of f(x) (call it $f^e(x)$) may be written as:

$$f^e(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} I(|\frac{x-x_i}{h}| < 1)$$

There is a connection with histograms, suppose no x_i lands exactly on the boundary of a bin, then this estimator corresponds to splitting the range of the random variable into bins of width 2h allowing x to be the "center" of each bin and treating $f^e(x)$ as the ordinate of the histogram. The

problem with this type of estimator is that it is not "smooth" but consists of a sequence of jumps at x ± h with a zero derivative everywhere else. Kernel estimators get around this problem by replacing .5I( ) in the above formula by a **kernel function** K( ) with certain desirable properties that to some degree resolve the "smoothness" problem. So that our estimator looks like:

$$f^e(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$$

Where h is usually referred to as the band width, window width or smoothing parameter.

The kernel function.

Generally a kernel function would be selected so that it satisfies:

$$\int_{-\infty}^{\infty} K(y)dy = 1.$$

making the vast array of continuous density functions suitable candidates. Provided K() is everywhere non negative (making it a density function) f^e( ) will itself be a density function and will inherit all of the continuity and differentiability properties of K.

Three Examples:

In each of the following h is the bandwidth and $t_i = (x-X_i)/h$.
The **Epanechnikov** Kernel $K_E(t_i)$ is of the form:

$$K_E(t_i) = \frac{.75(1-.2t_i^2)}{\sqrt{5}} \quad t_i^2 < 5$$
$$= 0 \quad otherwise$$

The **Gaussian** Kernel $K_G(t_i)$ is of the form:

$$K_G(t_i) = \frac{1}{(2\pi)^{.5}} e^{\frac{t_i^2}{2}}$$

The **Biweight** Kernel $K_B(t_i)$ is of the form:

$$K_B(t_i) = \frac{15}{16}(1-t^2)^2 \quad |t| < 1; \quad 0 \quad otherwise.$$

Note also the original naive estimator is like a rectangular kernel with $K(t_i) = .5$ $|t_i| < 1$, 0 otherwise.

Choosing the 'h' and the Kernel.

Think about the mean integrated squared error of the estimator defined by:

$$MISE(f^e) = E\int (f^e(x) - f(x))^2 dx$$

Since the integrand is non-negative the order of integration and expectation can be reversed, note also that $E(f^e - f)^2 = E(f^e - E(f^e) + E(f^e) - f)^2 = (E(f^e) - f)^2 + E(f^e - E(f^e))^2 = bias^2 + var(f^e)$ yielding:

$$MISE(f^e) = \int (E(f^e(x)) - f(x))^2 dx + \int var(f^e(x)) dx$$

which is the integrated squared bias plus the integrated variance. This would conceptually be a useful thing to minimize in choosing h and the Kernel. Rewriting K* = K/h it can be noted that:

$$E(f^e) = \frac{1}{n}\sum_{i=1}^{n} E(K^*(t_i)) = \int K^*(t) f(x) dx$$

which, for given f, does not depend upon n but only on K and h. This indicates that taking larger samples alone will not reduce the bias, attention has to be focused on the choice of h and K!

Confining attention to Kernels symmetric about zero with continuous derivatives at all orders with a variance v_k, it can be shown that (see Silverman pages 39 to 40) that the optimal h is equal to:

$$v_k^{-\frac{2}{5}}\left(\int K(t)^2 dt\right)^{\frac{1}{5}}\left(\int f''(x)^2 dx\right)^{\frac{-1}{5}} n^{\frac{-1}{5}}$$

Unfortunately "optimal h" here depends upon knowledge of the unknown f( ) we are attempting to estimate, however it does tell us that the optimal window gets smaller as the sample size grows (last term) and as the degree of fluctuation of the unknown function increases (penultimate term). Substituting the value of the optimal h back into the formula for the mean integrated squared error and minimizing with respect to K results in the Epanechnikov Kernel. The relative efficiencies of other kernels can be shown to be .9512 for the Gaussian kernel, .9939 for the Biweight kernel and .9295 for the rectangular suggesting that there is little to choose between kernels on efficiency grounds.

Choosing h.

Referring to the normal family of distributions with a variance σ^2, yields a value of "optimal h" of $1.06\ \sigma\ n^{-.2}$. One could then estimate σ from the data and, on the presumption that the distribution being estimated was like the normal, use this as the value for h. When the underlying distribution is not normal this tends to result in over-smoothing (especially when bi-modality is present). A safe alternative, based upon a sample standard deviation of σ and a sample inter-quartile range of ξ, the bandwidth 'h' is specified as:

$$h = \frac{.9\min\left(\sigma, \frac{\xi}{1.34}\right)}{n^{\frac{1}{5}}}$$

Least Squares Cross Validation.

Noting that the integrated squared error may be written as:

$$\int (f^e - f)^2 = \int f^{e2} - 2\int f^e f + \int f^2$$

and that the last term does not depend upon f^e and hence h interest focuses on minimizing an approximation to the first two terms with respect to h. After some tedious argument it may be shown that a good approximation to these two terms is:

$$\frac{1}{n^2 h}\sum_i \sum_j K^c\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0)$$

where K^c is defined as:

$$K^c(t) = K^2(t) - 2K(t)$$

and $K^2(t)$ is the convolution of the Kernel with itself. Numerical methods for minimizing this w.r.t. h can easily consume inordinate amounts of time however fourier transform methods can be used to substantially reduce computations (see Silverman P61-66). Non the less the computational burden remains considerable.

Likelihood Cross Validation

Let $f_{-i}(\ )$ be the Kernel estimate calculated by missing out observation x_i then $\ln f_{-1}(x_i)$ is the log-likelihood of f as the density underlying the independent additional observation x_i . We can think of maximizing this with respect to h, indeed why not maximize:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \ln f_{-i}(x_i)$$

This is related to the Kullback-Leibler Information distance $I(f,f^e)$ where:

$$I(f,f^e) = \int f(x)\ln\left(\frac{f(x)}{f^e(x)}\right)dx$$

Thinking of E(CV(h)) as the expectation of f_{-j}^{e} for some arbitrary j we have:

$$E(\ln f_{-j}^{e}) = \int f \ln f_{-j}^{e} dx = \int f(\ln f_{-j}^{e} - \ln f)dx + \int f \ln f dx \approx -I(f, f^{e}) + \int f \ln f dx$$

thus we are minimizing the Kullback - Leibler information distance plus a constant. Alternatively Silverman suggests eyeballing the problem. Plot out a selection of curves based upon different h's and choose the one that best suits ones priors.

A variable bandwidth h: The Adaptive Kernel.

One of the problems with the above estimators is that the degree of smoothing is constant over all x, the same value in regions densely populated with x as it is in regions sparsely populated with x. This can lead to over-smoothing in the dense areas (taking out "bumps" that should be there) and / or under-smoothing in sparse areas (leaving in bumps that should not be there). To solve this problem a variable bandwidth estimator has been developed. Essentially in this case our estimator f^{ae} is of the form:

$$f^{ae}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - x_i}{h_i}\right)$$

Where h_i is usually referred to as the local band width, window width or smoothing parameter. Various methods are available for estimation in this case, one of the simplest, most practical and most effective is the following. Compute f^{e} in one of the preceding methods which yields a fixed bandwidth h. Calculate f^{gm} the geometric mean of f^{e} for the sample given by:

$$f^{gm} = \prod_{i=1}^{n} f^{e}(X_i)^{\frac{1}{n}}$$

set h_i as:

$$h_i = h\left(\frac{f^{e}(X_i)}{f^{gm}}\right)^{-a}$$

where a is a sensitivity parameter chosen by the investigator. Generally $0 \leq a \leq 1$ with a=0 returning us to the fixed bandwidth estimator. Most applications seem to choose a = 0.5.

Consistency

Under apparently very mild conditions on the kernel namely,

$$\int |K(t)|\,dt < \infty, \qquad \int K(t)dt = 1.$$

together with $|tK(t)| \to 0$ as $|t| \to \infty$ and a window width h_n satisfying $h_n \to 0$ as $nh_n \to \infty$ convergence in probability of $f^e(x)$ to f(x) (convergence at a point) can be established. Essentially the requirement on h is that it does not converge to 0 as rapidly as n^{-1} ensuring the expected number of points in $x \pm h_n$ tends to infinity with n. Further, and more importantly, under similar conditions $\sup_x |f^e(x) - f(x)|$ can also be shown to converge to 0. A note of caution, the rate of convergence is often very slow so that in this instance consistency is by no means a warranty for good estimates!