10.0 Lesson Plan

- Answer Questions
- Robust Estimators
- Maximum Likelihood Estimators

10.1 Robust Estimators

Previously, we claimed to like estimators that are unbiased, have minimum variance, and/or have minimum mean squared error. Typically, one cannot achieve all of these properties with the same estimator.

An estimator may have good properties for one distribution, but not for another. We saw that $\frac{n}{n-1}Z$, for Z the sample maximum, was excellent in estimating θ for a Unif$(0, \theta)$ distribution. But it would not be excellent for estimating θ when, say, the density function looks like a triangle supported on $[0, \theta]$.

A **robust estimator** is one that works well across many families of distributions. In particular, it works well when there may be outliers in the data.

The **10% trimmed mean** is a robust estimator of the population mean. It discards the 5% largest and 5% smallest observations, and averages the rest. (Obviously, one could trim by some fraction other than 10%, but this is a commonly-used value.)

Surveyors distinguish errors from blunders. Errors are measurement jitter attributable to chance effects, and are approximately Gaussian. Blunders occur when the guy with the theodolite is standing on the wrong hill.

A trimmed mean throws out the blunders and averages the good data. If all the data are good, one has lost some sample size. But in exchange, you are protected from the corrosive effect of outliers.

10.2 Strategies for Finding Estimators

Some economists focus on the **method of moments**. This is a terrible procedure—its only virtue is that it is easy.

But easy virtue is always problematic. I don't think this topic is worth our attention.

Instead, we shall focus on

- maximum likelihood estimators
- Bayesian estimators.

10.3 Maximum Likelihood Estimators

Recall the function that links the probability of random variables to parameters:

$$f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m).$$

When the $x_1, \ldots, x_n$ are treated as variables and the the parameters $\theta_1, \ldots, \theta_m$ are treated as constants, this is the **joint density function**.

But when the $x_1, \ldots, x_n$ are treated as constants (the values observed in the sample) and the the $\theta_1, \ldots, \theta_m$ are treated as variables, this is the **likelihood function**.

The **maximum likelihood estimates** of the parameters $\theta_1, \ldots, \theta_m$ are the values $\hat{\theta}_1, \ldots, \hat{\theta}_m$ that maximize the likelihood function.

This procedure was invented by Sir Ronald Fisher when he was an undergraduate at Cambridge. His intuition was that one wants the estimate of the parameter values that gives the largest "probability" of having obtained the sample $X_1, \ldots, X_n$ that was actually observed, i.e., the $x_1, \ldots, x_n$.

Fisher was an astonishing statistician. He also invented the mathematical theory of population genetics, the theory of experimental design, and other areas.

In general, one can show that maximum likelihood estimators

- have bias that goes to zero for large sample sizes
- have approximately minimum variance for large sample sizes
- often have approximately normal distributions for large sample sizes.

Additionally, if one is interested in estimating some function $h(\theta)$ rather than the actual parameter itself, one can prove that the maximum likelihood estimate of $h(\theta)$ is $h(\hat{\theta})$. This is not generally true for unbiased estimators or minimum variance unbiased estimators.

Trick: When maximizing the likelihood function, it is often easier to maximize the log of the likelihood function. Taking the log converts the product to the sum. Since the log is a monotonic function, its maximum must occur for the same value of θ as does the likelihood function.

Example 1: Let $x_1, \ldots, x_n$ be an observed random sample from an exponential distribution with parameter λ. We want to find the maximum likelihood estimate of λ.

First, we find the likelihood function (i.e., the joint density of the data, where the sample values are known but the λ is not):

$$\begin{aligned} f(x_1, ..., x_n; \lambda) &= \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) \\ &= \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i). \end{aligned}$$

Next, we solve this to find the value of λ that maximizes the likelihood function. We could try to do this directly, but that leads to a difficult derivative. Instead, we use the **Trick**: the value of λ which maximizes $f(x_1, ..., x_n; \lambda)$ is the same value that maximizes $\ln f(x_1, ..., x_n; \lambda)$.

Let $\ell(\lambda) = \ln f(x_1, ..., x_n; \lambda)$. Then

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

and to maximize this we take the derivative with respect to λ, set it to 0, and solve:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

so the MLE of λ, denoted by a circumflex, is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\bar{x}.$$

Is this MLE biased?

$$\mathbb{E}[\hat{\lambda}] = \mathbb{E}[1/\bar{X}] \neq 1/\left(\mathbb{E}[\bar{X}]\right)$$

So it is a biased estimator. In fact, with more work, one can show $\mathbb{E}[\hat{\lambda}] = \frac{n}{n+1}\lambda$.

Example 2: Sometimes the trick of taking the log does not work. Consider again the problem of estimating θ in a Unif$(0, \theta)$ distribution.

The joint density function is

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &= \prod_{i=1}^{n} f(x_i; \theta) \qquad \textbf{Why?} \\ &= \prod_{i=1}^{n} \frac{1}{\theta} I(0 \leq x_i \leq \theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \leq x_i \leq \theta) \end{aligned}$$

where $I(0 \leq x_i \leq \theta)$ is an **indicator function** that takes the value 1 iff $0 \leq x_i \leq \theta$ and is zero otherwise.

The indicator function is a slick way to carry the information that the density is zero below 0 and above θ.

Let $z = \max\{x_1, \ldots x_n\}$. Then

$$\prod_{i=1}^{n} I(0 \le x_i \le \theta) = I(0 \le z \le \theta)$$

and so the likelihood function is

$$f(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n} I(0 \le z \le \theta).$$

We need to maximize this function with respect to θ. Clearly the function is 0 for all values of θ that are less than z. For values greater than z, as θ increases, $(1/\theta)^n$ gets smaller. So the maximum must occur at $\hat{\theta} = z$.

Thus the maximum likelihood estimate of θ is the sample maximum. From previous work we know this is slightly biased—recall that the bias is $-\theta/(n+1)$—but the bias goes to zero as the sample size increases.

Example 3:

Let $x_1, .., x_n$ be an observed random sample from a Normal distribution with unknown mean μ and unknown standard deviation σ.

First, we find the likelihood function:

$$\begin{aligned} f(x_1, \ldots, x_n; \mu\sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]. \end{aligned}$$

This is a case in which the trick of taking the logarithm is helpful:

$$\ell(x_1, \ldots, x_n; \mu, \sigma) = n \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Take partial derivatives to compute $\hat{\mu}$ and $\hat{\sigma}$ and solve

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} \\ 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma}. \end{aligned}$$

Specifically, for the mean, we see:

$$\begin{aligned} 0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \mu} \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ &= \sum_{i=1}^{n} (x_i - \mu) = \left(\sum_{i=1}^{n} x_i \right) - n\mu. \end{aligned}$$

and solving this for μ shows that the maximum likelihood estimate (denoted by the circumflex on the parameter) is $\hat{\mu} = \bar{x}$.

Now, to find the MLE for σ, we take the derivative of the log-likelihood with respect to σ, set it to 0, and solve:

$$\begin{aligned}
0 &= \frac{\partial \ell(x_1, \ldots, x_n; \mu, \sigma)}{\partial \sigma} \\
&= \frac{\partial}{\partial \sigma}\left[n \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= \frac{\partial}{\partial \sigma}\left[-n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2.
\end{aligned}$$

Note that in the penultimate step we used properties of the logarithm to simplify the first term. Specifically, we used the fact that

$$n \ln \frac{1}{\sqrt{2\pi}\sigma} = -n \ln \sqrt{2\pi}\sigma = -n \ln \sqrt{2\pi} - n \ln \sigma.$$

Thus

$$\frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \textbf{ so } \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

We do not know μ, but it turns out the joint maximization wrt both parameters occurs when we substitute in the MLE of μ, so

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

We cut some corners in this derivation:

- A full solution requires checking a condition on the second derivatives to ensure we are maximizing the log-likelihood instead of minimizing it or finding an inflection point;
- The joint minimization wrt to both parameters requires solving a set of simultaneous equations, which is why we can substitute $\bar{x}$ for μ in finding the MLE for σ.