14.30 Introduction to Statistical Methods in Economics
Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 18

Konrad Menzel

April 23, 2009

1 Properties of Estimators (continued)

1.1 Standard Error

Often we also want to make statements about the precision of the estimator - we can always state the value of the estimate, but how confident are we that it is actually close to the true parameter?

Definition 1 *The* standard error $\sigma(\hat{\theta})$ *of an estimate is the standard deviation (or estimated standard deviation) of the estimator,*

$$SE(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta}(X_1, \ldots, X_n))}$$

Should recall that an estimator is a function of the random variables, and therefore a random variable for which we can calculate expectation, variance and other moments.

Example 1 *The mean $\bar{X}_n$ of an i.i.d. sample $X_1, \ldots, X_n$ where $\mathrm{Var}(X_i) = \sigma^2$ has variance $\frac{\sigma^2}{n}$. Therefore, the standard error is*

$$SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

If we don't know σ^2, we calculate the estimated *standard error*

$$\hat{SE}(\bar{X}_n) = \frac{\hat{\sigma}}{\sqrt{n}}$$

The standard error is a way of comparing the precision of estimators, and we'd obviously favor the estimator which has the smaller variance/standard error.

Definition 2 *If $\hat{\theta}_A$ and $\hat{\theta}_B$ are unbiased estimators for θ, i.e. $\mathbb{E}_{\theta_0}[\hat{\theta}_A] = \mathbb{E}_{\theta_0}[\hat{\theta}_B] = \theta_0$, then we say that $\hat{\theta}_A$ is* efficient relative to $\hat{\theta}_B$ *if*

$$\mathrm{Var}(\hat{\theta}_B) \geq \mathrm{Var}(\hat{\theta}_A)$$

Sometimes we look at an entire *class* of estimators $\Theta = \{\hat{\theta}_1, \hat{\theta}_2, \ldots\}$, and we say that $\hat{\theta}_A$ is efficient in that class if it has the lowest variance of all members of Θ.

Example 2 *Suppose that X and Y are scores from two different Math tests. You are interested in some underlying "math ability", and the two scores are noisy (and possibly correlated) measurements with $\mathbb{E}[X] = \mathbb{E}[Y] = \mu$, $\mathrm{Var}(X) = \sigma_X^2$, $\mathrm{Var}(Y) = \sigma_Y^2$, and $\mathrm{Cov}(X, Y) = \sigma_{XY}$. Instead of using only one of the measurements, you decide to combine them into a weighted average $pX + (1-p)Y$ instead. What is the* *expectation of this weighted average? Which value of p minimizes the variance of the weighted average? We can interpret this as an estimation problem in which we want to estimate μ using a sample of only two observations. Since all weighted averages of X and Y have mean μ, we'll try to find the efficient estimator.*

From the formula of the variance of a sum of random variables,

$$\mathrm{Var}(pX + (1-p)Y) = p^2\sigma_X^2 + 2p(1-p)\sigma_{XY} + (1-p)^2\sigma_Y^2$$

In order to find the optimal p, we set the first derivative equal to zero, i.e.

$$0 = 2p\sigma_X^2 + 2(1-2p)\sigma_{XY} - 2(1-p)\sigma_Y^2$$

Solving for p, we get, assuming that $\sigma_X^2 + \sigma_Y^2 > 2\sigma_{XY}$ (notice that this is also the sufficient condition for a local minimum)

$$p^* = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 - 2\sigma_{XY} + \sigma_Y^2} = \frac{\mathrm{Var}(Y) - \mathrm{Cov}(X,Y)}{\mathrm{Var}(Y-X)} = \frac{\mathrm{Cov}(Y-X,Y)}{\mathrm{Var}(Y-X)}$$

Note that if X and Y are uncorrelated, the efficient estimator puts weight $p^* = \frac{\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}$ on X which is greater the lower the variance of X is relative to that of Y.

2 Methods for Constructing Estimators

2.1 Method of Moments

This method was proposed by the British statistician Karl Pearson in 1894: suppose we have to estimate k parameters of a distribution. then we can look at the first k *sample moments* of the data,

$$\begin{aligned}
\bar{X}_n &= \frac{1}{n}\sum_{i=1}^n X_i \\
\overline{X_n^2} &= \frac{1}{n}\sum_{i=1}^n X_i^2 \\
&\vdots \\
\overline{X_n^k} &= \frac{1}{n}\sum_{i=1}^n X_i^k
\end{aligned}$$

and equate them to the corresponding *population moments* for a given parameter value, calculated under the distribution,

$$\begin{aligned}
\mu_1(\theta) &= \mathbb{E}_\theta[X_i] \equiv \int_{-\infty}^{\infty} x f_X(x;\theta)dx \\
&\vdots \\
\mu_k(\theta) &= \mathbb{E}_\theta[X_i^k] \equiv \int_{-\infty}^{\infty} x^k f_X(x|\theta)dx
\end{aligned} \tag{1}$$

Then the method of moments (MoM) estimator $\hat{\theta}$ can be obtained by solving the equations

$$\mu_j(\hat{\theta}) = \overline{X_n^j} \qquad j = 1, \ldots, k$$

for θ.

Example 3 *Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from a Poisson distribution with unknown parameter λ, i.e. $X \sim P(\lambda)$. The distribution has only one unknown parameter, and the first population moment is given by*

$$\mu_1(\lambda) = \mathbb{E}_\lambda[X] = \lambda$$

Therefore, the MoM estimator is given by

$$\hat{\lambda} = \mu_1(\hat{\lambda}) \equiv \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

What if we used more moments than necessary to estimate the parameter? - We also know that for the Poisson distribution

$$\mathbb{E}_\lambda[X^2] = \mathrm{Var}_\lambda(X) + \mathbb{E}_\lambda[X]^2 = \lambda + \lambda^2$$

Example 4 *A* double exponential *random variable has p.d.f.*

$$f_Y(y) = \frac{1}{2}\lambda e^{-\lambda|y-\mu|}$$

so we have to estimate two parameters (λ, μ). We can look up in a statistics book that

$$\mathbb{E}[Y] = \mu \qquad \mathbb{E}[Y^2] = \mathrm{Var}(Y) + \mathbb{E}[Y]^2 = \frac{2}{\lambda^2} + \mu^2$$

so the method of moments estimator solves

$$\begin{aligned} \bar{Y} &= \hat{\mu} \\ \overline{Y^2} &= \frac{2}{\hat{\lambda}^2} + \hat{\mu}^2 \end{aligned}$$

so that, solving for $(\hat{\lambda}, \hat{\mu})$,

$$\hat{\mu} = \bar{Y}, \qquad \hat{\lambda} = \sqrt{2}\left(\overline{Y^2} - (\bar{Y})^2\right)^{-1/2}$$

2.2 Maximum Likelihood Estimation

While the method of moments only tries to match a selected number of moments of the population to their sample counterparts, we might alternatively construct an estimator which makes the population distribution as a whole match the sample distribution as closely as possible. This is what the *maximum likelihood estimator* of a parameter θ does, which is loosely speaking, the value which "most likely" would have generated the observed sample:

Suppose we have an i.i.d. sample $Y_1, \ldots, Y_n$ where the p.d.f. of Y is given by $f_Y(y|\theta)$, which is known up to the parameter θ. The Maximum Likelihood estimator (MLE) is a function $\hat{\theta}$ of the data maximizing the *joint p.d.f.* of the data under θ.

More specifically, we define the *likelihood* of the sample as

$$\mathcal{L}(\theta) = f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Usually it is much easier to maximize the logarithm of the likelihood function,

$$L(\theta) = \log(\mathcal{L}(\theta)) = \sum_{i=1}^{n} \log f(y_i|\theta)$$

Note that since the logarithm is a strictly increasing function, $\mathcal{L}(\theta)$ and $L(\theta)$ will be maximized at the same value.

Proposition 1 *The expectation of the log-likelihood under the parameter θ_0,*

$$\mathbb{E}_{\theta_0}[L(\theta)] = \mathbb{E}[\log f(Y|\theta)]$$

is maximized at the true parameter θ_0.

Proof: Since the true density over which we take the expectation is $f(y|\theta_0)$, we can show that $\mathbb{E}_{\theta_0}[L(Y|\theta) - L(Y|\theta_0)] \leq 0$ for all values of θ using Jensen's Inequality and the fact that $\log(\cdot)$ is concave

$$\begin{aligned}
\mathbb{E}_{\theta_0}[L(Y|\theta) - L(Y|\theta_0)] &= \mathbb{E}_{\theta_0}[\log f(Y|\theta) - \log f(Y|\theta_0)] = \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(Y|\theta)}{f(Y|\theta_0)}\right)\right] \\
&\leq \log\left(\mathbb{E}_{\theta_0}\left[\frac{f(y|\theta)}{f(y|\theta_0)}\right]\right) = \log\left(\int_{-\infty}^{\infty} \frac{f(y|\theta)}{f(y|\theta_0)} f(y|\theta_0)dy\right) \\
&= \log\left(\int_{-\infty}^{\infty} f(y|\theta)dy\right) = \log(1) = 0
\end{aligned}$$

since $f(y|\theta)$ is a density and therefore integrates to 1. Therefore $\mathbb{E}_{\theta_0}[L(Y|\theta_0)] \geq \mathbb{E}_{\theta_0}[L(Y|\theta)]$ for all values of θ, so that θ_0 maximizes the function $\square$

Since by the Law of Large Numbers, the log likelihood for and i.i.d. sample

$$\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\theta) \xrightarrow{p} \mathbb{E}[\log f(Y|\theta)]$$

we'd think that maximizing the log likelihood for a large i.i.d. sample should therefore give us a parameter "close" to θ_0.

Example 5 *Suppose $X \sim N(\mu_0, \sigma_0^2)$, and we want to estimate the parameters μ and σ^2 from an i.i.d. sample $X_1, \ldots, X_n$. The likelihood function is*

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

It turns out that it's much easier to maximize the log-likelihood,

$$\begin{aligned}
\log \mathcal{L}(\theta) &= \sum_{i=1}^{n} \log\left\{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}\right\} \\
&= \sum_{i=1}^{n}\left\{\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(X_i-\mu)^2}{2\sigma^2}\right\} \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2
\end{aligned}$$

In order to find the maximum, we take the derivatives with respect to μ and σ^2, and set them equal to zero:

$$0 = \frac{1}{2\widehat{\sigma^2}}\sum_{i=1}^{n} 2(X_i - \hat{\mu}) \Leftrightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Similarly,

$$0 = -\frac{n}{2}\frac{2\pi}{2\pi\widehat{\sigma^2}} + \frac{1}{2\left(\widehat{\sigma^2}\right)^2}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \Leftrightarrow \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Recall that we already showed that this estimator is not unbiased for σ_0^2, so in general, Maximum Likelihood Estimators need not be unbiased.

Example 6 *Going back to the example with the uniform distribution, suppose $X \sim U[0, \theta]$, and we are interested in estimating θ. For the method of moments estimator, you can see that*

$$\mu_1(\theta) = \mathbb{E}_\theta[X] = \frac{\theta}{2}$$

so equating this with the sample mean, we obtain

$$\hat{\theta}_{MoM} = 2\bar{X}_n$$

What is the maximum likelihood estimator? Clearly, we wouldn't pick any $\hat{\theta} \leq \max\{X_1, \ldots, X_n\}$ because a sample with realizations greater than $\hat{\theta}$ has zero probability under $\hat{\theta}$. Formally, the likelihood is

$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } 0 \leq X_i \leq \theta \text{ for all } i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

We can see that any value of $\theta \leq \max\{X_1, \ldots, X_n\}$ can't be a maximum because $L(\theta)$ is zero for all those points. Also, for $\theta \geq \max\{X_1, \ldots, X_n\}$ the likelihood function is strictly decreasing in θ, and therefore, it is maximized at

$$\hat{\theta}_{MLE} = \max\{X_1, \ldots, X_n\}$$

Note that since $X_i < \theta_0$ with probability 1, the Maximum Likelihood estimator is also going to be less than θ_0 with probability one, so it's not unbiased. More specifically, the p.d.f. of $X_{(n)}$ is given by

$$f_{X_{(n)}}(y) = n[F_X(y)]^{n-1}f_X(y) = \begin{cases} \frac{n}{\theta_0}\left(\frac{1}{\theta_0}\frac{y}{\theta_0}\right)^{n-1} & \text{if } 0 \leq y \leq \theta_0 \\ 0 & \text{otherwise} \end{cases}$$

so that

$$\mathbb{E}[X_{(n)}] = \int_{-\infty}^{\infty} y f_{X_{(n)}}(y)dy = \int_0^{\theta_0} n\left(\frac{y}{\theta_0}\right)^n dy = \frac{n}{n+1}\theta_0$$

We could easily construct an unbiased estimator $\tilde{\theta} = \frac{n+1}{n}X_{(n)}$.

2.3 Properties of the MLE

The following is just a summary of main theoretical results on MLE (won't do proofs for now)

- If there is an efficient estimator in the class of consistent estimators, MLE will produce it.

- Under certain regularity conditions, MLE's will have an asymptotically normal distribution (this comes essentially from an application of the Central Limit Theorem)

Is Maximum Likelihood always the best thing to do? - not necessarily

- may be biased
- often hard to compute
- might be sensitive to incorrect assumptions on underlying distribution