4. Method of moments: LM 5.2, P. 357-8; Jan 13, 2010

4.1 One-parameter method of moments

Suppose we have an n-sample, $Y_1, \ ..., \ Y_n$ from some probability distribution, indexed by a parameter θ. Suppose $\mathrm{E}(Y) = g(\theta)$. Then to estimate θ we solve $g(\theta) = (1/n)\sum_i Y_i \equiv \overline{Y_n}$.

If $g(\theta) \propto \theta$, we may get a reasonable estimator:

For example: $Y_i \sim \mathcal{P}o(\theta)$; $\mathrm{E}(Y_i) = \theta$; estimator is $\overline{Y_n}$.

or $Y \sim B(r, \theta)$; $\mathrm{E}(Y_i) = r\theta$; estimator is $\overline{Y_n}/r = (1/nr)\sum_i Y_i$.

Even if $g(\theta) \propto \theta$, we may not get a good estimator:

For example: $Y_i \sim U(0, \theta)$; $\mathrm{E}(Y_i) = \theta/2$; estimator is $2\overline{Y_n}$, but we know estimator based on $\max_i(Y_i)$ is "better".

Another example: $Y_i \sim \mathcal{E}(\lambda)$; $\mathrm{E}(Y_i) = 1/\lambda$; estimator of λ is $1/\overline{Y_n}$; we will come back to this one.

4.2 k-parameter method of moments

If we have k parameters, $\theta_1,....,\theta_k$, and $\mathrm{E}(Y) = g_1(\theta_1, ..., \theta_k)$, $\mathrm{E}(Y^2) = g_2(\theta_1, ..., \theta_k)$, ... $\mathrm{E}(Y^k) = g_k(\theta_1, ..., \theta_k)$, Then, given sample values $y_1, ..., y_n$, to get the estimates we solve

$$\begin{aligned} \frac{1}{n}\sum_i y_i &= g_1(\theta_1, ..., \theta_k) \\ \frac{1}{n}\sum_i y_i^2 &= g_2(\theta_1, ..., \theta_k) \\ &... \\ \frac{1}{n}\sum_i y_i^k &= g_k(\theta_1, ..., \theta_k) \end{aligned}$$

In general we use as many equations as we need to get solutions for our parameters.

4.3 An example where we need to use Y^2 for one parameter

Suppose $Y_1, ..., Y_n$ are an n-sample from $N(0, \sigma^2)$

$\mathrm{E}(Y_i) = 0$; first equation is no use.

$\mathrm{E}(Y_i^2) = \sigma^2$; estimator of σ^2 is $(1/n)\sum_i Y_i^2$.

But suppose we wanted an estimator of σ: solving the equation we would obtain MoM estimator $\sqrt{(1/n)\sum_i Y_i^2}$. Is this a "good" estimator? We will return to this one later.

4.4 An example with two parameters

Suppose $Y_1, ..., Y_n$ are an n-sample from $N(\mu, \sigma^2)$

Then we have the two equations $\mathrm{E}(Y_i) = \mu$; $\mathrm{E}(Y_i^2) = \mu^2 + \sigma^2$ (why?)

Estimator of μ is $\overline{Y_n}$; this seems reasonable – estimate the population mean by the sample mean.

Estimator of σ^2 is

$$\begin{aligned} (1/n)\sum_i Y_i^2 - (\overline{Y_n})^2 &= (1/n)(\sum_i Y_i^2 - n\overline{Y_n}^2) = (1/n)(\sum_i (Y_i - \overline{Y_n})^2) \\ \text{since } \sum_i (Y_i - \overline{Y_n})^2 &= \sum_i Y_i^2 - 2\overline{Y_n}\sum_i Y_i + n\overline{Y_n}^2 = \sum_i Y_i^2 - 2.\overline{Y_n}.n\overline{Y_n} + n\overline{Y_n}^2 \end{aligned}$$

This might seem reasonable: mean squared deviation about the population mean, estimated by mean squared deviation of the sample about the sample mean.

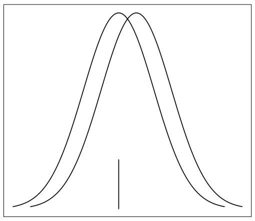
However, this is not the estimate of variance your calculator typically calculates.

We define $S^2 = \sum_i (Y_i - \overline{Y_n})^2$.

5. Principles of Estimation: estimator T(X) ; Jan 15, 2010; part 1

Here we consider data random variables $\mathbf{X}$, which may or may nor be an n-sample $X_1, ..., X_n$. The estimator T is some function of the data random variables.

5.1 Unbiasedness



Bias: $b_T(\theta) \;=\; \mathrm{E}_\theta(T) - \theta$

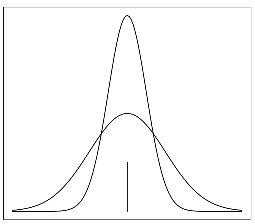
Estimator is unbiased if $b_T(\theta) \;=\; 0$ for all θ

For example: $X \;\sim Bin(n,\theta)$: $\mathrm{E}(X/n) \;=\; \theta$.

In this example, the estimator X/n is unbiased for θ.

Other things being equal, we would like to have unbiased estimators.

5.2 Variance

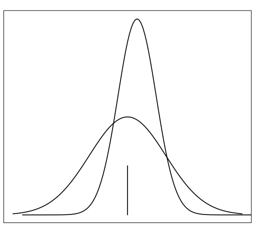


Unbiased estimators with small variance have higher probability of providing estimates close to the true value.

$\mathrm{var}_\theta(T) \;=\; \mathrm{E}_\theta(T - \mathrm{E}(T))^2 \;=\; \mathrm{E}_\theta(T-\theta)^2$ if $\mathrm{E}_\theta(T) = \theta$.

Among unbiased estimators, we prefer those with small variance.

5.3 Mean square error

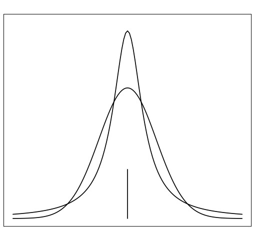


There is a trade-off between bias and variance: an estimator with non-zero bias but small variance may have smaller mean square error.

$\mathrm{mse}_\theta(T) \;=\; \mathrm{E}_\theta(T-\theta)^2 \;=\; \mathrm{E}_\theta((T - \mathrm{E}_\theta(T)) + (\mathrm{E}_\theta(T) - \theta))^2)$

$= \;\mathrm{var}_\theta(T) + 2.b_T(\theta).\mathrm{E}_\theta(T - \mathrm{E}_\theta(T)) \;+\; (b_T(\theta))^2 \;=\; \mathrm{var}_\theta(T) + (b_T(\theta))^2$

5.4 Mean square error or mean absolute error?



Squared error gives high weight to large probabilities: For example, a 1 in 10^6 probability of an error of 10^6 gives a contribution $10^{-6} \times (10^6)^2 \;=\; 10^6$ contribution to the mse.

We might be more interested in mean absolute error:

that is $\mathrm{E}_\theta(|T-\theta|)$.

With this measure, a 1 in 10^6 probability of an error of 10^6 gives a contribution $10^{-6} \times (10^6) \;=\; 1$ contribution to the mean absolute error.

6. Examples of estimators and their properties; Jan 15, 2010; part 2

6.1 Sample mean is unbiased estimator of population mean

(i) If $Y_1, ..., Y_n$ are i.i.d $\sim \mathcal{P}o(\theta)$, MoM estimator of θ is $\overline{Y_n} = \sum_{i=1}^n Y_i/n$, and is unbiased for θ.

(ii) If $Y_1, ..., Y_n$ are i.i.d $\sim Bin(r, \theta)$, MoM estimator of θ is $\sum_{i=1}^n Y_i/(rn)$, and is unbiased for θ.

(iii) If $Y_1, ..., Y_n$ are i.i.d $\sim N(\mu, \sigma^2)$, MoM estimator of μ is $\overline{Y_n} = \sum_{i=1}^n Y_i/n$, and is unbiased for μ, regardless of the value of σ^2, and of whether or not we are also estimating σ^2.

6.2 Estimation of σ^2 and of σ

Consider again $Y_1, ..., Y_n$ i.i.d $\sim N(0, \sigma^2)$, (see 4.3). The MoM estimator of σ^2 is $T = (1/n)\sum_{i=1}^n Y_i^2$.
The estimator T is unbiased for σ^2, since $\mathrm{E}(Y_i^2) = \sigma^2$. Bias $b_T(\sigma^2) = \mathrm{E}(T) - \sigma^2 = 0$.
The MoM estimator of σ is $\sqrt{T}$.
We know for any (non-degenerate) Y; $0 < \mathrm{var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2$, or $\mathrm{E}(Y^2) > (\mathrm{E}(Y))^2$.
So $\sigma^2 = \mathrm{E}(T) > (\mathrm{E}(\sqrt{T}))^2$; i.e. $\mathrm{E}(\sqrt{T}) < \sigma$.
The MoM estimator of σ is biased; the bias $b_{\sqrt{T}}(\sigma) = \mathrm{E}(\sqrt{T}) - \sigma < 0$.

6.3 Estimation of θ from an exponential sample

Suppose $Y_1, ..., Y_n$ are i.i.d $\sim \mathcal{E}(\theta)$: $f_Y(y) = \theta\exp(-\theta y)$ on $0 \le y < \infty$ (0 otherwise).
$\mathrm{E}(Y) = 1/\theta$; so MoM estimator of θ is $T = 1/\overline{Y_n}$.
We know $\mathrm{E}(\overline{Y_n}) = 1/\theta$; what can we say about $\mathrm{E}(T) = \mathrm{E}(1/\overline{Y_n})$??
When V and W are independent, $\mathrm{E}(VW) = \mathrm{E}(V).\mathrm{E}(W)$.
If V and W tends to be large/small together, $\mathrm{E}(VW) > \mathrm{E}(V).\mathrm{E}(W)$.
If V tends to be large when W is small, and vice versa, $\mathrm{E}(VW) < \mathrm{E}(V).\mathrm{E}(W)$.
Now $1/\overline{Y_n}$ tends to be large when $\overline{Y_n}$ is small, and vice versa.
Thus $1 = \mathrm{E}(\overline{Y_n}.(1/\overline{Y_n})) < \mathrm{E}(\overline{Y_n}).\mathrm{E}(T)$ or $\mathrm{E}(T) > 1/\mathrm{E}(\overline{Y_n}) = 1/(1/\theta) = \theta$.
That is MoM estimator T is positively biased: $b_T(\theta) = \mathrm{E}(T) - \theta > 0$.

6.4 Estimation of σ^2 when μ is unknown

$$\begin{aligned}\sum_i (Y_i - \mu)^2 &= \sum_i ((Y_i - \overline{Y_n}) + (\overline{Y_n} - \mu))^2 \\ &= S^2 + (\overline{Y_n} - \mu)\sum_i (Y_i - \overline{Y_n}) + n.(\overline{Y_n} - \mu)^2 = S^2 + n(\overline{Y_n} - \mu)^2\end{aligned}$$

Now $\mathrm{E}((Y_i - \mu)^2) = \sigma^2$ (why?) and $\mathrm{E}((\overline{Y_n} - \mu)^2) = \sigma^2/n$ (why?) so $\mathrm{E}(S^2) = n.\sigma^2 - n.\sigma^2/n = (n-1)\sigma^2$.
For an unbiased estimator, we would use $S^2/(n-1)$ (like most calculators), not the MoM S^2/n.

6.5 Sample from $U(0, \theta)$

$\mathrm{var}(Y_i) = \theta^2/12$; variance of the MoM estimator is $\theta^2/(3n)$. It is unbiased.
What is the variance of $W = \max(Y_i)$? $f_W(w) = nw^{n-1}/\theta^n$ on $0 < w < \theta$.
$\mathrm{E}(W) = n\theta/(n+1)$. $\mathrm{E}(W^2) = n\theta^2/(n+2)$. So $\mathrm{var}(W) = n\theta^2(1/(n+2) - n/(n+1)^2) = n\theta^2/(n+2)(n+1)^2$.
Or $\mathrm{var}((n+1)W/n) = \theta^2/n(n+2)$. This estimator is unbiased, and, if $n > 1$, has smaller variance than the MoM estimator.

Consider the estimator $KW(n+1)/n$; bias is $(K-1)\theta$; variance is $K^2\theta^2/n(n+2)$.
MSE is $\theta^2(K^2/n(n+2) + (K-1)^2)$ minimized by $K = n(n+2)/(n+1)^2$.
That is, estimator $(n+2)W/(n+1)$ has smallest MSE.