Mean Square Error (MSE)

Introduction

Let $X_1, X_2, \cdots, X_n$ be n i.i.d. random variables, i.e., a random sample from $f(x|\theta)$, where θ is unknown. An estimator of θ is a function of (only) the n random variables, i.e., a statistic $\hat{\theta} = r(X_1, \cdots, X_n)$. There are several method to obtain an estimator for θ, such as the MLE, method of moment, and Bayesian method.

A difficulty that arises is that since we can usually apply more than one of these methods in a particular situation, we are often face with the task of choosing between estimators. Of course, it is possible that different methods of finding estimators will yield the same answer (as we have see in the MLE handout), which makes the evaluation a bit easier, but, in many cases, different methods will lead to different estimators. We need, therefore, some criteria to choose among them.

Mean Square Error (MSE) of an Estimator

Let $\hat{\theta}$ be the estimator of the unknown parameter θ from the random sample $X_1, X_2, \cdots, X_n$. Then clearly the deviation from $\hat{\theta}$ to the true value of θ, $|\hat{\theta} - \theta|$, measures the quality of the estimator, or equivalently, we can use $(\hat{\theta} - \theta)^2$ for the ease of computation. Since $\hat{\theta}$ is a random variable, we should take average to evaluation the quality of the estimator. Thus, we introduce the following

Definition: The mean square error (MSE) of an estimator $\hat{\theta}$ of a parameter θ is the function of θ defined by $E(\hat{\theta} - \theta)^2$, and this is denoted as $MSE_{\hat{\theta}}$.

This is also called the risk function of an estimator, with $(\hat{\theta} - \theta)^2$ called the quadratic loss function. The expectation is with respect to the random variables $X_1, \cdots, X_n$ since they are the only random components in the expression.

Notice that the MSE measures the average squared difference between the estimator $\hat{\theta}$ and the parameter θ, a somewhat reasonable measure of performance for an estimator. In general, any increasing function of the absolute distance $|\hat{\theta} - \theta|$ would serve to measure the goodness

of an estimator (mean absolute error, $E(|\hat{\theta} - \theta|)$, is a reasonable alternative. But MSE has at least two advantages over other distance measures: First, it is analytically tractable and, secondly, it has the interpretation

$$MSE_{\hat{\theta}} = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 = Var(\hat{\theta}) + (Bias\ of\ \hat{\theta})^2$$

This is so because

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E(\hat{\theta}^2) + E(\theta^2) - 2\theta E(\hat{\theta}) \\ &= Var(\hat{\theta}) + [E(\hat{\theta})]^2 + \theta^2 - 2\theta E(\hat{\theta}) \\ &= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \end{aligned}$$

Thus, MSE has two components, one measures the variability of the estimator (precision) and the other measures the its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias. To find an estimator with good MSE properties, we need to find estimators that control both variance and bias.

For an unbiased estimator $\hat{\theta}$, we have

$$MSE_{\hat{\theta}} = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta})$$

and so, if an estimator is unbiased, its MSE is equal to its variance.

Example 1: Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. random variables with density functi $f(x|\sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right)$, the maximum likelihood estimator for σ

$$\hat{\sigma} = \frac{\sum_{i=1}^{n} |X_i|}{n}$$

is unbiased.

Solution: Let us first calculate $E(|X|)$ and $E(|X|^2)$ as

$$\begin{aligned} E(|X|) &= \int_{-\infty}^{\infty} |x| f(x|\sigma) dx = \int_{-\infty}^{\infty} |x| \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right) dx \\ &= \sigma \int_0^{\infty} \frac{x}{\sigma} \exp\left(-\frac{x}{\sigma}\right) d\frac{x}{\sigma} = \sigma \int_0^{\infty} y e^{-y} dy = \sigma \Gamma(2) = \sigma \end{aligned}$$

and

$$\begin{aligned} E(|X|^2) &= \int_{-\infty}^{\infty} |x|^2 f(x|\sigma) dx = \int_{-\infty}^{\infty} |x|^2 \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right) dx \\ &= \sigma^2 \int_0^{\infty} \frac{x^2}{\sigma^2} \exp\left(-\frac{x}{\sigma}\right) d\frac{x}{\sigma} = \sigma^2 \int_0^{\infty} y^2 e^{-y} dy = \sigma \Gamma(3) = 2\sigma^2 \end{aligned}$$

Therefore,

$$E(\hat{\sigma}) = E\left(\frac{|X_1| + \cdots + |X_n|}{n}\right) = \frac{E(|X_1|) + \cdots + E(|X_n|)}{n} = \sigma$$

So $\hat{\sigma}$ is an unbiased estimator for σ.

Thus the MSE of $\hat{\sigma}$ is equal to its variance, i.e.

$$\begin{aligned}
MSE_{\hat{\sigma}} &= E(\hat{\sigma} - \sigma)^2 = Var(\hat{\sigma}) = Var\left(\frac{|X_1| + \cdots + |X_n|}{n}\right) \\
&= \frac{Var(|X_1|) + \cdots + Var(|X_n|)}{n^2} = \frac{Var(|X|)}{n} \\
&= \frac{E(|X|^2) - (E(|X|))^2}{n} = \frac{2\sigma^2 - \sigma^2}{n} = \frac{\sigma^2}{n}
\end{aligned}$$

The Statistic S^2: Recall that if $X_1, \cdots, X_n$ come from a normal distribution with variance σ^2, then the sample variance S^2 is defined as

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

It can be shown that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. From the properties of χ^2 distribution, we have

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1 \Rightarrow E(S^2) = \sigma^2$$

and

$$Var\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1) \Rightarrow Var(S^2) = \frac{2\sigma^4}{n-1}$$

Example 2: Let $X_1, X_2, \cdots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with expected value μ and varianc
σ^2, then $\bar{X}$ is an unbiased estimator for μ, and S^2 is an unbiased estimator for σ^2.

Solution: We have

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{E(X_1) + \cdots + E(X_n)}{n} = \mu$$

Therefore, $\bar{X}$ is an unbiased estimator. The MSE of $\bar{X}$ is

$$MSE_{\bar{X}} = E(\bar{X} - \mu)^2 = Var(\bar{X}) = \frac{\sigma^2}{n}$$

This is because

$$Var(\bar{X}) = Var\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{Var(X_1) + \cdots + Var(X_n)}{n^2} = \frac{\sigma^2}{n}$$

Similarly, as we showed above, $E(S^2) = \sigma^2$, S^2 is an unbiased estimator for σ^2, and the MSE of S^2 is given by

$$MSE_{S^2} = E(S^2 - \sigma^2) = Var(S^2) = \frac{2\sigma^4}{n-1}.$$

Although many unbiased estimators are also reasonable from the standpoint of MSE, be aware that controlling bias does not guarantee that MSE is controlled. In particular, it is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in MSE.

Example 3: An alternative estimator for σ^2 of a normal population is the maximum likelihood or method of moment estimator

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$$

It is straightforward to calculate

$$E(\hat{\sigma^2}) = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2$$

so $\hat{\sigma^2}$ is a biased estimator for σ^2. The variance of $\hat{\sigma^2}$ can also be calculated as

$$Var(\hat{\sigma^2}) = Var\left(\frac{n-1}{n}S^2\right) = \frac{(n-1)^2}{n^2}Var(S^2) = \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}.$$

Hence the MSE of $\hat{\sigma}^2$ is given by

$$\begin{aligned} E(\hat{\sigma}^2 - \sigma^2)^2 &= Var(\hat{\sigma}^2) + (Bias)^2 \\ &= \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

We thus have (using the conclusion from Example 2)

$$MSE_{\hat{\sigma}^2} = \frac{2n-1}{n^2}\sigma^4 < \frac{2n}{n^2}\sigma^4 = \frac{2\sigma^4}{n} < \frac{2\sigma^4}{n-1} = MSE_{S^2}.$$

This shows that $\hat{\sigma}^2$ has smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved.

The above example does not imply that S^2 should be abandoned as an estimator of σ^2. The above argument shows that, on average, $\hat{\sigma}^2$ will be closer to σ^2 than S^2 if MSE is used as a measure. However, $\hat{\sigma}^2$ is biased and will, on the average, underestimate σ^2. This fact alone may make us uncomfortable about using $\hat{\sigma}^2$ as an estimator for σ^2.

In general, since MSE is a function of the parameter, there will not be one "best" estimator in terms of MSE. Often, the MSE of two estimators will cross each other, that is, for some parameter values, one is better, for other values, the other is better. However, even this partial information can sometimes provide guidelines for choosing between estimators.

One way to make the problem of finding a "best" estimator tractable is to limit the class of estimators. A popular way of restricting the class of estimators, is to consider only unbiased estimators and choose the estimator with the lowest variance.

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of a parameter θ, that is, $E(\hat{\theta}_1) = \theta$ and $E(\hat{\theta}_2) = \theta$, then their mean squared errors are equal to their variances, so we should choose the estimator with the smallest variance.

Example 4: This problem is connected with the estimation of the variance of a normal distribution with unknown mean from a sample $X_1, X_2, \cdots, X_n$ of i.i.d. normal random variables. For what value of ρ does $\rho \sum_{i=1}^{n}(X_i - \bar{X})^2$ have the minimal MSE?

Please note that if $\rho = \frac{1}{n-1}$, we get S^2 in example 2; when $\rho = \frac{1}{n}$, we get $\hat{\sigma}^2$ in example 3.

Solution:

As in above examples, we define

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Then,

$$E(S^2) = \sigma^2 \quad \text{and} \quad \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

Let

$$e_\rho = \rho \sum_{i=1}^{n} (X_i - \bar{X})^2 = \rho(n-1)S^2$$

and let $t = \rho(n-1)$ Then

$$E(e_\rho) = \rho(n-1)E(S^2) = \rho(n-1)\sigma^2 = t\sigma^2$$

and

$$Var(e_\rho) = \rho^2(n-1)^2 Var(S^2) = \frac{2t^2}{n-1}\sigma^4$$

We can Calculate the MSE of e_ρ as

$$\begin{aligned} MSE(e_\rho) &= Var(e_\rho) + [Bias]^2 = Var(e_\rho) + \left[E(e_\rho) - \sigma^2\right]^2 \\ &= Var(e_\rho) + (t\sigma^2 - \sigma^2)^2 = Var(e_\rho) + (t-1)^2\sigma^4. \end{aligned}$$

Plug in the results before, we have

$$MSE(e_\rho) = \frac{2t^2}{n-1}\sigma^4 + (t-1)^2\sigma^4 = f(t)\sigma^4$$

where

$$f(t) = \frac{2t^2}{n-1} + (t-1)^2 = \left(\frac{n+1}{n-1}t^2 - 2t + 1\right)$$

when $t = \frac{n-1}{n+1}$, $f(t)$ achieves its minimal value, which is $\frac{2}{n+1}$. That is the minimal value of $MSE(e_\rho) = \frac{2\sigma^4}{n+1}$, with $(n-1)\rho = t = \frac{n-1}{n+1}$, i.e. $\rho = \frac{1}{n+1}$.

From the conclusion in example 3, we have

$$MSE_{\hat{\sigma^2}} = \frac{2n-1}{n^2}\sigma^4 < \frac{2\sigma^4}{n-1} = MSE_{S^2}.$$

It is straightforward to verify that

$$MSE_{\hat{\sigma^2}} = \frac{2n-1}{n^2}\sigma^4 \geq \frac{2\sigma^4}{n+1} = MSE(e_\rho)$$

when $\rho = \frac{1}{n+1}$.

Exercise 1. X, the cosine of the angle at which electrons are emitted in muon decay has a density

$$f(x) = \frac{1+\alpha X}{2} \quad -1 \leq x \leq 1 \quad -1 \leq \alpha \leq 1$$

The parameter α is related to polarization. Show that $E(X) = \frac{\alpha}{3}$. Consider an estimator for the parameter α, $\hat{\alpha} = 3\bar{X}$. Computer the variance, the bias, and the mean square error of this estimator.