Mean squared error

From Wikipedia, the free encyclopedia

In statistics, the **mean squared error** (**MSE**) or **mean squared deviation** (**MSD**) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.[1]

The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated. In an analogy to standard deviation, taking the square root of MSE yields the root-mean-square error or root-mean-square deviation (RMSE or RMSD), which has the same units as the quantity being estimated; for an unbiased estimator, the RMSE is the square root of the variance, known as the standard deviation.

Contents

Definition and basic properties

The MSE assesses the quality of an **estimator** (i.e., a mathematical function mapping a sample of data to a parameter of the population from which the data is sampled) or a **predictor** (i.e., a function mapping arbitrary inputs to a sample of values of some random variable). Definition of an MSE differs according to whether one is describing an estimator or a predictor.

Predictor

If $\hat{\boldsymbol{Y}}$ is a vector of $\boldsymbol{n}$ predictions, and $\boldsymbol{Y}$ is the vector of observed values corresponding to the inputs to the function which generated the predictions, then the MSE of the predictor can be estimated by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

I.e., the MSE is the *mean* $\left(\frac{1}{n}\sum_{i=1}^{n}\right)$ of the *square of the errors* ($(\hat{Y}_i - Y_i)^2$). This is an easily computable quantity for a particular sample (and hence is sample-dependent).

Estimator

The MSE of an estimator $\hat{\theta}$ with respect to an unknown parameter θ is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right].$$

This definition depends on the unknown parameter, and the MSE in this sense is a property of an estimator. Since an MSE is an expectation, it is not technically a random variable. That being said, the MSE could be a function of unknown parameters, in which case any *estimator* of the MSE based on estimates of these parameters would be a function of the data and thus a random variable. If the estimator is derived from a sample statistic and is used to estimate some population statistic, then the expectation is with respect to the sampling distribution of the sample statistic.

The MSE can be written as the sum of the variance of the estimator and the squared bias of the estimator, providing a useful way to calculate the MSE and implying that in the case of unbiased estimators, the MSE and variance are equivalent.[2]

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2.$$

Proof of variance and bias relationship

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2 + 2\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)\left(\mathbb{E}[\hat{\theta}] - \theta\right) + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] + \mathbb{E}\left[2\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)\left(\mathbb{E}[\hat{\theta}] - \theta\right)\right] + \mathbb{E}\left[\left(\mathbb{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] + 2\left(\mathbb{E}[\hat{\theta}] - \theta\right)\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right] + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 && \mathbb{E}[\hat{\theta}] - \theta = \text{const.} \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] + 2\left(\mathbb{E}[\hat{\theta}] - \theta\right)\left(\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}]\right) + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 && \mathbb{E}[\hat{\theta}] = \text{const.} \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 \\
&= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2
\end{aligned}$$

Regression

In regression analysis, the term *mean squared error* is sometimes used to refer to the unbiased estimate of error variance: the residual sum of squares divided by the number of degrees of freedom. This definition for a known, computed quantity differs from the above definition for the computed MSE of a predictor in that a different denominator is used. The denominator is the sample size reduced by the number of model parameters estimated from the same data, *(n-p)* for *p* regressors or *(n-p-1)* if an intercept is used.[3] For more details, see errors and residuals in statistics. Note that, although the MSE (as defined in the present article) is not an unbiased estimator of the error variance, it is consistent, given the consistency of the predictor.

Also in regression analysis, "mean squared error", often referred to as mean squared prediction error or "out-of-sample mean squared error", can refer to the mean value of the squared deviations of the predictions from the true values, over an out-of-sample test space, generated by a model estimated over a particular sample space. This also is a known, computed quantity, and it varies by sample and by out-of-sample test space.

Examples

Mean

Suppose we have a random sample of size *n* from a population, $X_1, \ldots, X_n$. Suppose the sample units were chosen with replacement. That is, the *n* units are selected one at a time, and previously selected units are still eligible for selection for all *n* draws. The usual estimator for the mean is the sample average

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

which has an expected value equal to the true mean μ (so it is unbiased) and a mean square error of

$$\operatorname{MSE}\left(\overline{X}\right) = \operatorname{E}\left[\left(\overline{X} - \mu\right)^2\right] = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \frac{\sigma^2}{n}$$

where σ^2 is the population variance.

For a Gaussian distribution this is the best unbiased estimator (that is, it has the lowest MSE among all unbiased estimators), but not, say, for a uniform distribution.

Variance

The usual estimator for the variance is the *corrected sample variance:*

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

This is unbiased (its expected value is σ^2), hence also called the *unbiased sample variance,* and its MSE is[4]

$$\operatorname{MSE}(S_{n-1}^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{1}{n}\left(\gamma_2 + \frac{2n}{n-1}\right)\sigma^4,$$

where μ_4 is the fourth central moment of the distribution or population and $\gamma_2 = \mu_4/\sigma^4 - 3$ is the excess kurtosis.

However, one can use other estimators for σ^2 which are proportional to S_{n-1}^2, and an appropriate choice can always give a lower mean square error. If we define

$$S_a^2 = \frac{n-1}{a} S_{n-1}^2 = \frac{1}{a} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

then we calculate:

$$\begin{aligned}
\operatorname{MSE}(S_a^2) &= \mathbb{E}\left[\left(\frac{n-1}{a} S_{n-1}^2 - \sigma^2 \right)^2 \right] \\
&= \mathbb{E}\left[\frac{(n-1)^2}{a^2} S_{n-1}^4 - 2\left(\frac{n-1}{a} S_{n-1}^2 \right) \sigma^2 + \sigma^4 \right] \\
&= \frac{(n-1)^2}{a^2} \mathbb{E}\left[S_{n-1}^4 \right] - 2\left(\frac{n-1}{a} \right) \mathbb{E}\left[S_{n-1}^2 \right] \sigma^2 + \sigma^4 \\
&= \frac{(n-1)^2}{a^2} \mathbb{E}\left[S_{n-1}^4 \right] - 2\left(\frac{n-1}{a} \right) \sigma^4 + \sigma^4 && \mathbb{E}\left[S_{n-1}^2 \right] = \sigma^2 \\
&= \frac{(n-1)^2}{a^2} \left(\frac{\gamma_2}{n} + \frac{n+1}{n-1} \right) \sigma^4 - 2\left(\frac{n-1}{a} \right) \sigma^4 + \sigma^4 && \mathbb{E}\left[S_{n-1}^4 \right] = \operatorname{MSE}(S_{n-1}^2) + \sigma^4 \\
&= \frac{n-1}{na^2} \left((n-1)\gamma_2 + n^2 + n \right) \sigma^4 - 2\left(\frac{n-1}{a} \right) \sigma^4 + \sigma^4
\end{aligned}$$

This is minimized when

$$a = \frac{(n-1)\gamma_2 + n^2 + n}{n} = n + 1 + \frac{n-1}{n}\gamma_2.$$

For a Gaussian distribution, where $\gamma_2 = 0$, this means the MSE is minimized when dividing the sum by $a = n + 1$. The minimum excess kurtosis is $\gamma_2 = -2$,[a] which is achieved by a Bernoulli distribution with *p* = 1/2 (a coin flip), and the MSE is minimized for $a = n - 1 + \frac{2}{n}$. So no matter what the kurtosis, we get a "better" estimate (in the sense of having a lower MSE) by scaling down the unbiased estimator a little bit; this is a simple example of a shrinkage estimator: one "shrinks" the estimator towards zero (scales down the unbiased estimator).

Further, while the corrected sample variance is the best unbiased estimator (minimum mean square error among unbiased estimators) of variance for Gaussian distributions, if the distribution is not Gaussian then even among unbiased estimators, the best unbiased estimator of the variance may not be S_{n-1}^2.

Gaussian distribution

The following table gives several estimators of the true parameters of the population, μ and σ^2, for the Gaussian case.[5]

True value	Estimator	Mean squared error
$\theta = \mu$	$\hat{\theta}$ = the unbiased estimator of the population mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i)$	$\mathrm{MSE}(\overline{X}) = \mathrm{E}((\overline{X} - \mu)^2) = \left(\frac{\sigma}{\sqrt{n}}\right)^2$
$\theta = \sigma^2$	$\hat{\theta}$ = the unbiased estimator of the population variance, $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$	$\mathrm{MSE}(S_{n-1}^2) = \mathrm{E}((S_{n-1}^2 - \sigma^2)^2) = \frac{2}{n-1}\sigma^4$
$\theta = \sigma^2$	$\hat{\theta}$ = the biased estimator of the population variance, $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$	$\mathrm{MSE}(S_n^2) = \mathrm{E}((S_n^2 - \sigma^2)^2) = \frac{2n-1}{n^2}\sigma^4$
$\theta = \sigma^2$	$\hat{\theta}$ = the biased estimator of the population variance, $S_{n+1}^2 = \frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$	$\mathrm{MSE}(S_{n+1}^2) = \mathrm{E}((S_{n+1}^2 - \sigma^2)^2) = \frac{2}{n+1}\sigma^4$

Note that:

1. The MSEs shown for the variance estimators assume $X_i \sim \mathrm{N}($ n (as measured by MSE): the MSE of S_{n-1}^2 is larger than that of S_{n+1}^2 or S_n^2.
2. Estimators with the smallest total variation may produce biased estimates: S_{n+1}^2 typically underestimates σ^2 by $\frac{2}{n}\sigma^2$

Interpretation

An MSE of zero, meaning that the estimator $\hat{\theta}$ predicts observations of the parameter θ with perfect accuracy, is the ideal, but is typically not possible.

Values of MSE may be used for comparative purposes. Two or more statistical models may be compared using their MSEs as a measure of how well they explain a given set of observations: An unbiased estimator (estimated from a statistical model) with the smallest variance among all unbiased estimators is the best unbiased estimator or MVUE (Minimum Variance Unbiased Estimator).

Both linear regression techniques such as analysis of variance estimate the MSE as part of the analysis and use the estimated MSE to determine the statistical significance of the factors or predictors under study. The goal of experimental design is to construct experiments in such a way that when the observations are analyzed, the MSE is close to zero relative to the magnitude of at least one of the estimated treatment effects.

MSE is also used in several stepwise regression techniques as part of the determination as to how many predictors from a candidate set to include in a model for a given set of observations.

Applications

- Minimizing MSE is a key criterion in selecting estimators: see minimum mean-square error. Among unbiased estimators, minimizing the MSE is equivalent to minimizing the variance, and the estimator that does this is the minimum variance unbiased estimator. However, a biased estimator may have lower MSE; see estimator bias.

- In statistical modelling the MSE, representing the difference between the actual observations and the observation values predicted by the model, is used to determine the extent to which the model fits the data and whether the removal or some explanatory variables, simplifying the model, is possible without significantly harming the model's predictive ability.

Loss function

Squared error loss is one of the most widely used loss functions in statistics, though its widespread use stems more from mathematical convenience than considerations of actual loss in applications. Carl Friedrich Gauss, who introduced the use of mean squared error, was aware of its arbitrariness and was in agreement with objections to it on these grounds.[1] The mathematical benefits of mean squared error are particularly evident in its use at analyzing the performance of linear regression, as it allows one to partition the variation in a dataset into variation explained by the model and variation explained by randomness.

Criticism

The use of mean squared error without question has been criticized by the decision theorist James Berger. Mean squared error is the negative of the expected value of one specific utility function, the quadratic utility function, which may not be the appropriate utility function to use under a given set of circumstances. There are, however, some scenarios where mean squared error can serve as a good approximation to a loss function occurring naturally in an application.[6]

Like variance, mean squared error has the disadvantage of heavily weighting outliers.[7] This is a result of the squaring of each term, which effectively weights large errors more heavily than small ones. This property, undesirable in many applications, has led researchers to use alternatives such as the mean absolute error, or those based on the median.

See also

- James–Stein estimator
- Hodges' estimator
- Mean percentage error
- Mean square weighted deviation
- Mean squared displacement
- Mean squared prediction error
- Minimum mean squared error estimator
- Mean square quantization error
- Peak signal-to-noise ratio
- Root mean square deviation
- Squared deviations

Notes

a. This can be proved by Jensen's inequality as follows. The fourth central moment is an upper bound for the square of variance, so that the least value for their ratio is one, therefore, the least value for the excess kurtosis is −2, achieved, for instance, by a Bernoulli with p=1/2.

References

1. Lehmann, E. L.; Casella, George (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer. ISBN 0-387-98502-6. MR 1639875 (https://www.ams.org/mathscinet-getitem?mr=1639875).
2. Wackerly, Dennis; Mendenhall, William; Scheaffer, Richard L. (2008). *Mathematical Statistics with Applications* (7 ed.). Belmont, CA, USA: Thomson Higher Education. ISBN 0-495-38508-5.

3. Steel, R.G.D, and Torrie, J. H., *Principles and Procedures of Statistics with Special Reference to the Biological Sciences.*, McGraw Hill, 1960, page 288.
4. Mood, A.; Graybill, F.; Boes, D. (1974). *Introduction to the Theory of Statistics* (3rd ed.). McGraw-Hill. p. 229.
5. DeGroot, Morris H. (1980). *Probability and Statistics* (2nd ed.). Addison-Wesley.
6. Berger, James O. (1985). "2.4.2 Certain Standard Loss Functions". *Statistical decision theory and Bayesian Analysis* (2nd ed.). New York: Springer-Verlag. p. 60. ISBN 0-387-96098-8. MR 0804611 (https://www.ams.org/mathscinet-getitem?mr=0804611).
7. Sergio Bermejo, Joan Cabestany (2001) "Oriented principal component analysis for large margin classifiers (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T08-43PS3GC-1&_user=483692&_coverDate=12%2F31%2F2001&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000022720&_version=1&_urlVersion=0&_userid=483692&md5=8586e409694e1b50da3aa3c6fce18cb8)", *Neural Networks*, 14 (10), 1447–1461.

Categories: Point estimation performance | Statistical deviation and dispersion | Loss functions | Least squares

---