Notes On Estimation

Alon Jacobson

August 5, 2021

1 Notation

Say we have a fixed, unknown quantity of interest θ which is some statistic or functional of an underlying statistical model. We gather scalar or vector-valued data which is generated from this statistical model, and this data that we obtain is described by a random variable X. From a realization of X, we construct a point estimate or "best guess" of θ by $\widehat{\theta} = g(X)$ which is some fixed function g of X. Thus $\widehat{\theta}$ is a random variable.

The bias of an estimator is defined by

$$\mathsf{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta = \mathbb{E}(\widehat{\theta} - \theta).$$

We say that $\widehat{\theta}$ is *unbiased* when $\mathsf{bias}(\widehat{\theta}) = 0$, or equivalently when $\mathbb{E}(\widehat{\theta}) = \theta$. We use $\mathbb{E}(\widehat{\theta})$ to mean the expected value over the distribution of data X (i.e., averaging over all possible observations of X): $\mathbb{E}(r(X)) = \int r(x) f_X(x)\, dx$. Similarly, we write $\mathbb{V}$ for the variance over the distribution of X.

A way to assess the quality of a point estimate is by the mean squared error, or MSE, defined by:

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

The MSE can be written as

$$\mathrm{MSE}(\widehat{\theta}) = \mathsf{bias}^2(\widehat{\theta}) + \mathbb{V}(\widehat{\theta}).$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathbb{E}[(\widehat{\theta} - \theta)^2] \\
&= (\mathbb{E}[\widehat{\theta} - \theta])^2 + \mathbb{V}(\widehat{\theta} - \theta) \\
&= \mathsf{bias}^2(\widehat{\theta}) + \mathbb{V}(\widehat{\theta}).
\end{aligned}$$

where we have used the fact that, since for a random variable X, $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$, we have $\mathbb{E}(X^2) = (\mathbb{E}X)^2 + \mathbb{V}(X)$. □

2 Bias Correction

Suppose you have a point estimate $\widehat{\theta}$ of a quantity θ and you want to correct for its bias to make it unbiased. If you know the bias, you could make a new estimator $\widehat{\theta}_1$:

$$\widehat{\theta}_1 := \widehat{\theta} - \mathsf{bias}(\widehat{\theta}).$$

This estimator is unbiased:

$$\begin{aligned}\mathsf{bias}(\widehat{\theta}_1) &= \mathbb{E}(\widehat{\theta}_1 - \theta) \\ &= \mathbb{E}(\widehat{\theta} - \mathsf{bias}(\widehat{\theta}) - \theta) \\ &= \mathbb{E}(\widehat{\theta} - \theta) - \mathbb{E}(\mathsf{bias}(\widehat{\theta})) \\ &= \mathsf{bias}(\widehat{\theta}) - \mathbb{E}(\mathsf{bias}(\widehat{\theta})) \\ &= \mathsf{bias}(\widehat{\theta}) - \mathsf{bias}(\widehat{\theta}) = 0.\end{aligned}$$

Now, suppose instead that you know that $\mathsf{bias}(\widehat{\theta}) = K\theta$ for some constant K. (Or, alternatively, you only estimate K with $\widehat{K}$, so you have $\mathsf{bias}(\widehat{\theta}) = K\theta \approx \widehat{K}\theta$. But this will not affect the calculations.) But of course you don't know θ itself, so you don't know $\mathsf{bias}(\widehat{\theta}) = K\theta$. So you plug in $\widehat{\theta}$ for θ to estimate $\mathsf{bias}(\widehat{\theta})$:

$$\widehat{\mathsf{bias}}(\widehat{\theta}) := K\widehat{\theta}.$$

So you construct the estimator

$$\begin{aligned}\widehat{\theta}_2 &:= \widehat{\theta} - \widehat{\mathsf{bias}}(\widehat{\theta}) \\ &= \widehat{\theta} - K\widehat{\theta} = (1-K)\widehat{\theta}.\end{aligned}$$

Is this estimator unbiased? Let's check.

$\widehat{\theta}_2$ is an estimator of the form $\alpha\widehat{\theta}$ for some constant α; in this case $\alpha = 1 - K$. Let's first calculate the bias of an estimator of this more general form (take note - this formula will be used a lot):

$$\begin{aligned}\mathsf{bias}(\alpha\widehat{\theta}) &= \mathbb{E}(\alpha\widehat{\theta}) - \theta \\ &= \alpha\mathbb{E}(\widehat{\theta}) - \alpha\theta - (1-\alpha)\theta \\ &= \alpha(\mathbb{E}(\widehat{\theta}) - \theta) - (1-\alpha)\theta \\ &= \alpha\mathsf{bias}(\widehat{\theta}) - (1-\alpha)\theta.\end{aligned}$$

Now we can compute $\mathsf{bias}(\widehat{\theta}_2)$:

$$\begin{aligned}\mathsf{bias}(\widehat{\theta}_2) &= \mathsf{bias}((1-K)\widehat{\theta}) \\ &= (1-K)\mathsf{bias}(\widehat{\theta}) - (1-(1-K))\theta \\ &= \mathsf{bias}(\widehat{\theta}) - K\mathsf{bias}(\widehat{\theta}) - K\theta \\ &= \mathsf{bias}(\widehat{\theta}) - K\mathsf{bias}(\widehat{\theta}) - \mathsf{bias}(\widehat{\theta}) \\ &= -K\mathsf{bias}(\widehat{\theta}) = -K^2\theta.\end{aligned}$$

So, assuming $K \neq 0$ and $\theta \neq 0$, $\widehat{\theta}_2$ is biased!

OK, let's take a different approach. We seek an estimator $\widehat{\theta}_3$ that is unbiased. Let's directly solve for $\widehat{\theta}_3$. No matter the value of $\widehat{\theta}_3$, as long as $\widehat{\theta} \neq 0$, we have

$$\widehat{\theta}_3 = \alpha\widehat{\theta}$$

for some $\alpha \in \mathbb{R}$. Let's solve for α by setting $\mathsf{bias}(\widehat{\theta}_3) = 0$.

$$\begin{aligned}\mathsf{bias}(\widehat{\theta}_3) = \mathsf{bias}(\alpha\widehat{\theta}) &= \alpha\mathsf{bias}(\widehat{\theta}) - (1-\alpha)\theta \\ &= \alpha K\theta - (1-\alpha)\theta \\ &= (\alpha K - (1-\alpha))\theta \\ &= ((1+K)\alpha - 1)\theta\end{aligned}$$

Set this to equal 0 and solve for α:

$$\begin{aligned}((1+K)\alpha - 1)\theta &= 0\\ (1+K)\alpha - 1 &= 0\\ \alpha &= \frac{1}{1+K}\end{aligned}$$

where we assumed that $\theta \neq 0$ and $K \neq -1$. So the unbiased estimator is

$$\widehat{\theta}_3 = \frac{1}{1+K}\widehat{\theta}.$$

So instead of multiplying by $1-K$ to get an unbiased estimate, we should divide by $1+K$. However, for K near 0, $\widehat{\theta}_2 = (1-K)\widehat{\theta}$ is almost unbiased, since $\mathsf{bias}(\widehat{\theta}_2) = -K^2\theta$ which gets small in magnitude relative to θ for K near 0. In fact, $\widehat{\theta}_2$ is a local linear approximation to $\widehat{\theta}_3$ about $K = 0$. One way to see this analytically is to write the Maclaurin series for $\frac{1}{1+K}$ using the well-known geometric series formula:

$$\frac{1}{1+K} = \frac{1}{1-(-K)} = \sum_{n=0}^{\infty}(-K)^n = \sum_{n=0}^{\infty}(-1)^n K^n = 1 - K + K^2 - K^3 + K^4 - \cdots$$

Therefore, for K near 0, $1 - K \approx \frac{1}{1+K}$ and hence $\widehat{\theta}_2 \approx \widehat{\theta}_3$. This means that if someone derives a bias-correction estimator of the form of $\widehat{\theta}_2$, and when calculating the value of the $\widehat{\theta}_2$, K is generally small in magnitude, they might not notice that the estimator is actually slightly unbiased.

3 Minimum MSE

But what if, rather than an unbiased estimator, you want an estimator with minimum mean squared error? Loosely speaking, you want an estimator that is as close as possible on average to the true value. Call this minimum MSE estimator $\widehat{\theta}_4$, and again write this estimator in the form

$$\widehat{\theta}_4 = \alpha\widehat{\theta}.$$

We will choose α such that $\widehat{\theta}_4$ achieves minimum mean squared error. Let's calculate:

$$\begin{aligned}\mathrm{MSE}(\widehat{\theta}_4) &= \mathsf{bias}^2(\widehat{\theta}_4) + \mathbb{V}(\widehat{\theta}_4)\\ &= (\mathsf{bias}(\alpha\widehat{\theta}))^2 + \mathbb{V}(\alpha\widehat{\theta})\\ &= [((1+K)\alpha - 1)\theta]^2 + \alpha^2\mathbb{V}(\widehat{\theta})\\ &= \theta^2((1+K)\alpha - 1)^2 + \mathbb{V}(\widehat{\theta})\alpha^2.\end{aligned}$$

We're treating $\mathrm{MSE}(\widehat{\theta}_4)$ as a function of α and seeking to find α that minimizes it. So let's find the derivative and set it to 0.

$$\begin{aligned}\frac{d\,\mathrm{MSE}(\widehat{\theta}_4)}{d\alpha} &= 2\theta^2((1+K)\alpha - 1)(1+K) + 2\mathbb{V}(\widehat{\theta})\alpha\\ &= 2\theta^2(1+K)^2\alpha - 2\theta^2(1+K) + 2\mathbb{V}(\widehat{\theta})\alpha\\ &= [2\theta^2(1+K)^2 + 2\mathbb{V}(\widehat{\theta})]\alpha - 2\theta^2(1+K) = 0\end{aligned}$$

Solving for α yields

$$\alpha^* = \frac{\theta^2(1+K)}{\theta^2(1+K)^2 + \mathbb{V}(\widehat{\theta})} = \frac{1}{1+K+\frac{\mathbb{V}(\widehat{\theta})}{\theta^2(1+K)}}.$$

But we need to check that $\text{MSE}(\widehat{\theta}_4)$ is a global minimum at the critical point α^*. Applying the second derivative test,

$$\frac{d^2\,\text{MSE}(\widehat{\theta}_4)}{d\alpha^2} = 2\theta^2(1+K)^2 + 2\mathbb{V}(\widehat{\theta})$$

which must be positive, as long as $K \neq -1$ and $\theta \neq 0$. In particular the second derivative is positive at α^*, so by the second-derivative test, the critical point is at a local minimum. Also, since the second derivative is always positive, $\text{MSE}(\widehat{\theta}_4)$ is a convex (parabola) function of α, hence the critical point is the global minimum.

Notice the form of α^*; it has θ in it, but θ is the very thing we are trying to estimate! If we estimate θ with $\widehat{\theta}$ in α (compared to other proposed methods, this was empirically seen to show the best results by running simulations), we get the following estimator:

$$\widehat{\theta}_4 := \frac{\widehat{\theta}}{1+K+\frac{\mathbb{V}(\widehat{\theta})}{\widehat{\theta}^2(1+K)}}.$$

Hence we must know $\mathbb{V}(\theta)$, or at least an approximation of it $\widehat{\mathbb{V}}(\widehat{\theta})$, to get a minimum MSE estimator.

Comparing the unbiased estimator $\widehat{\theta}_3$ and the minimum MSE estimator $\widehat{\theta}_4$, we see that $\widehat{\theta}_3$ is just $\widehat{\theta}_4$ but where $\mathbb{V}(\widehat{\theta})$ is set to be 0. Also, since $\widehat{\theta}_4$ is just $\widehat{\theta}_3$ but with an extra positive term added in the denominator, we see that $\widehat{\theta}_4 \leq \widehat{\theta}_3$, and hence $\widehat{\theta}_4$ should be negatively biased, i.e., tends to under-estimate θ on average. Directly calculating $\mathsf{bias}(\widehat{\theta}_4)$, and going back to the original formula for α^* by treating the $\widehat{\theta}$ in the denominator of $\widehat{\theta}_4$ (which was initially θ, and is now a plug-in estimator for θ) as fixed at θ, and setting $C = \frac{\mathbb{V}(\widehat{\theta})}{\theta^2}$,

$$\begin{aligned}
\mathsf{bias}\left(\frac{\widehat{\theta}}{1+K+\frac{C}{1+K}}\right) &= \frac{1}{1+K+\frac{C}{1+K}}\mathsf{bias}(\widehat{\theta}) - \left(1 - \frac{1}{1+K+\frac{C}{1+K}}\right)\theta \\
&= \frac{K\theta}{1+K+\frac{C}{1+K}} - \frac{(K+\frac{C}{1+K})\theta}{1+K+\frac{C}{1+K}} \\
&= \frac{-\frac{C}{1+K}}{1+K+\frac{C}{1+K}}\theta \\
&= -\frac{C}{1+2K+K^2+C}\theta \\
&= -\frac{C}{(1+K)^2+C}\theta = -\frac{\theta}{\frac{(1+K)^2}{C}+1}.
\end{aligned}$$

Hence we see that $\mathsf{bias}(\widehat{\theta}_4) \leq 0$. As the value of C increases, the magnitude of this bias will increase.

4 Bias Correction Is Maximum Likelihood

Now what if we instead try to solve the problem using maximum likelihood? We will do this like so: From the equation

$$\mathsf{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta = K\theta$$

and solving for $\mathbb{E}(\widehat{\theta})$ we get

$$\mathbb{E}(\widehat{\theta}) = (1+K)\theta.$$

And suppose, as in the minimum MSE part, that we know, or at least can approximate, $\mathbb{V}(\widehat{\theta})$. Write $\mu = \mathbb{E}(\widehat{\theta})$ and $\sigma^2 = \mathbb{V}(\widehat{\theta})$, and assume that $\widehat{\theta} \sim \mathcal{N}(\mu, \sigma^2)$. The likelihood function is then

$$\mathcal{L}(\theta) = f_{\widehat{\theta}}(\widehat{\theta}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\widehat{\theta}-\mu)^2}{2\sigma^2}\right).$$

So the log-likelihood is

$$\ell(\theta) = \log \mathcal{L}(\theta) = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(\widehat{\theta}-\mu)^2}{2\sigma^2}.$$

Differentiate and set the derivative to 0:

$$\frac{d\ell(\theta)}{d\theta} = -\frac{(\widehat{\theta}-\mu)(-(1+K))}{\sigma^2} = \frac{(\widehat{\theta}-(1+K)\theta)(1+K)}{\sigma^2} = 0$$

Solve for θ:

$$\theta = \frac{\widehat{\theta}}{1+K} = \widehat{\theta}_3.$$

Let's verify that this is the maximum using the second derivative test:

$$\frac{d^2\ell(\theta)}{d\theta^2} = -\frac{(1+K)^2}{\sigma^2}$$

which is negative as long as $K \neq -1$. So we find that our unbiased estimator is also the MLE assuming a normal distribution.

5 Variance and Standard Error

The fact that the unbiased estimator is also the MLE makes this estimator more appealing as MLE is very popular. It is also appealing compared to the minimum-MSE estimator because it does not require an estimation of the variance of $\widehat{\theta}$. But if we do have an estimation of the variance of $\widehat{\theta}$, then the variance of $\widehat{\theta}_3$ is

$$\widehat{\mathbb{V}}(\widehat{\theta}_3) = \widehat{\mathbb{V}}\left(\frac{\widehat{\theta}}{1+K}\right) = \frac{\widehat{\mathbb{V}}(\widehat{\theta})}{(1+K)^2}.$$

So we can get a standard error:

$$\widehat{\mathsf{se}}(\widehat{\theta}_3) = \frac{\widehat{\mathsf{se}}(\widehat{\theta})}{|1+K|}.$$

6 Conclusion

In conclusion, we found three "correction" estimators for the value of θ if we start with an estimation $\widehat{\theta}$ and know that $\mathsf{bias}(\widehat{\theta}) = K\theta$ for some number K. Assuming that $\theta \neq 0$ and $K \neq -1$, we find three estimators:

$$\widehat{\theta}_2 = (1-K)\widehat{\theta}, \quad \widehat{\theta}_3 = \frac{\widehat{\theta}}{1+K}, \quad \widehat{\theta}_4 = \frac{\widehat{\theta}}{1+K+\frac{\widehat{\mathbb{V}}(\widehat{\theta})}{\widehat{\theta}^2(1+K)}}.$$

$\widehat{\theta}_3$ corrects the bias of $\widehat{\theta}$ by multiplying by the factor $1/(1+K)$, and is also the MLE; $\widehat{\theta}_2$ is a naive bias correction that is actually slightly biased but very close to $\widehat{\theta}_3$ for K near 0; and $\widehat{\theta}_4$ attempts to minimize the mean squared error and requires an estimate of the variance of $\widehat{\theta}$, $\widehat{\mathbb{V}}(\widehat{\theta})$. $\widehat{\theta}_4$ is negatively biased, meaning it on average under-estimates θ. $\widehat{\theta}_3$ is $\widehat{\theta}_4$ but with $\widehat{\mathbb{V}}(\widehat{\theta}) = 0$.