Lectures 18-21

4 Criteria for estimators

[CB7.3, BD3.4]

Definition 1 *The bias of an estimate $T(X)$ of a parameter $q(\theta)$ in a model (non-empty set of pdf/pmf) mathcalP $= P_\theta : \theta \in \Theta$ as $Bias_\theta(T) = E_\theta(T(X)) - q(\theta)$. An estimate such that $Bias_\theta(T) = 0$ is called unbiased. Any function $q(\theta)$ for which an unbiased estimate T exists is called an estimable function.*

This notion has intuitive appeal, ruling out, for instance, estimates that ignore the data, such as $T(X) = q(\theta_0)$, which can't be beat for $\theta = \theta_0$ but can obviously be arbitrarily terrible.
Eg: $\bar{X}$ and s^2 in normal distribution are unbiased for μ and σ^2. However, note that S is not an unbiased estimate of σ. Eg: (Unbiased estimates may be absurd) Let $X \sim Poisson(\lambda)$ and let $q(\lambda) = e^{-2\lambda}$. Consider $T(X) = (-1)^X$ as an estimate. It is unbiased but since T alternates between -1 and 1 while $q(\lambda) > 0$, it is not a good estimate.
Eg: (Unbiased Estimates in Survey Sampling) Suppose we wish to sample from a finite population, for instance, a census unit, to determine the average value of a variable (say) monthly family income during a time between two censuses and suppose that we have available a list of families in the unit with family incomes at the last census. Write $x_l, \cdots, x_N$ for the unknown current family incomes and correspondingly $u_1, \cdots, u_N$ for the known last census incomes. We ignore difficulties such as families moving. We let $X_l, \cdots, X_n$ denote the incomes of a sample of n families drawn at random without replacement. The parameter of interest is $\frac{1}{N}\sum_{i=1}^N x_i$. The model is

$$P(X_1 = a_1, \cdots X_n = a_n) = \binom{N}{n}^{-1} \quad \text{if} \quad \{a_1, \cdots, a_n\} \subseteq \{x_1, \cdots, x_n\}$$

Ex: $\bar{X}$ is unbiased and has variance $\frac{\sigma^2}{n}(1 - \frac{(n-1)}{(N-1)})$ where $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^2$. This method of sampling does not use the information contained in $u_1, \cdots, u_N$. One way to do this, reflecting the probable correlation between $(u_1, \cdots, u_N)$ and $(x_l, \cdots, x_N)$, is to estimate by a regression estimate

$$\bar{X}_R = \bar{X} - b(\bar{(U)} - \bar{u})$$

Ex: For each b this is unbiased.
Ex: If the correlation between U_i and X_i is positive (population) and $b < 2Cov(\bar{U}, \bar{X})/Var(\bar{U})$, this is better than $\bar{X}$.
Ex: The optimal choice of b is $Cov(\bar{U}, \bar{X})/Var(\bar{U})$. This value is unknown and can be estimated by

$$b_{opt} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(U_i - \bar{U})}{\frac{1}{N}\sum_{i=1}^N (u_i - \bar{u})^2}$$

. Ex: This estimator is biased.

4.1 Uniform Minimum Variance Unbiased (UMVU)

Note: If there exist 2 unbiased estimates T_1 and T_2 of θ, then any estimate of the form $\alpha T_1 + (1-\alpha)T_2$ for $0 \leq \alpha \leq 1$ will also be an unbiased estimate of θ. Which one should we choose?
For unbiased estimates mean square error and variance coincide.

Definition 2 *An unbiased estimate $T*(X)$ of $q(\theta)$ that has minimum MSE among all unbiased estimates for all θ is called UMVU (uniformly minimum variance unbiased). If this happens for a single parameter value θ_0 then it is locally minimum variance unbiased.*

Theorem 1 *Let U be the class of all unbiased estimates T of $\theta \in \Theta$ with $E_\theta(T^2) < \infty \forall \theta$, and suppose that U is non–empty. Let U_0 be the set of all unbiased estimates of 0, i.e., $U_0 = \{\nu : E_\theta(\nu) = 0, E_\theta(\nu^2) < \infty \forall \theta \in \Theta\}$. Then $T_0 \in U$ is UMVUE iff $E_\theta(\nu T_0) = 0 \forall \theta \in \Theta \forall \nu \in U_0$.*

Eg: Let X be unif$(\theta, \theta+1)$. Then $T = X - 1/2$ is unbiased for θ. An unbiased estimator $\nu(X)$ of zero has to satisfy $\int_\theta^{\theta+1} \nu(x)dx = 0$ for all θ. One such function is $\nu(x) = sin(2\pi x)$.

$$Cov(X - 1/2, sin(2\pi X)) = -cos(2\pi\theta)/2\pi.$$

This is non-zero. So T is not UMVU.
Eg: $X_1, \cdots, X_n$ iid unif(0,θ). Here $Y = (n+1)T/n$ is unbiased with T as $X_{(n)}$. Note that T is a sufficient statistic. We need to check if this is uncorrelated with all unbiased estimators of zero. Suppose W is an unbiased estimator of zero and cov(W,Y)>0. Then cov(E(W—Y),Y)=E(YE(W—Y))-E(W)E(Y)=EE(WY—Y)-E(W)E(Y)=cov(W,Y)¿0. So wlog, W can be considered a function of Y, equivalently a function of T. But T is complete sufficient implying W=0. Since Y is uncorrelated with W, Y is UMVE.

Theorem 2 *Let U be the non–empty class of unbiased estimates of $\theta \in \Theta$ as defined in Theorem 1. Then there exists at most one UMVUE $T \in U$ for θ.*

Theorem 3 *(Rao–Blackwell) Let W be any unbiased estimator of $\tau(\theta)$ and T be a sufficient statistic for θ. Define $\phi(T) = E(W \mid T)$. Then $\phi(T)$ is an estimator with $E(\phi(T)) = \tau(\theta)$ and $var(\phi(T)) \leq var(W)$.*

Pf: CB pf 342
This process of conditioning an unbiased estimator on a sufficient statistic is called Rao Blackwellization and leads to another unbiased estimator with uniformly lower variance. In other words, it is enough to consider the class of

unbiased estimators that are functions of sufficient statistics as any other unbiased estimator will have higher variance than one of them (the corresponding conditional correlation).

Eg: Suppose that $X_1, \cdots, X_n$ comes from density $\lambda exp(-\lambda x)$. Suppose that we want an estimate of $\theta = exp(-10\lambda)$. This corresponds to the probability P[$X_i > 10$]. The maximum likelihood estimate of λ is $1/\bar{X}$, so we could certainly claim $T = exp(-10/\bar{X})$ is the MLE of θ. This is certainly not unbiased.
Use statistic $u(X) = I(X_1 > 10)$. This statistic takes only the values 0 and 1, and it only depends on the first observation, so it's certainly a bad estimate. It is, however, unbiased.
The Rao-Blackwell theorem says that we can get a better unbiased estimate by using $u^*(X) = E[u(X)|V]$ where $V = \sum X_i$ is a sufficient statistic.
The conditional distribution of X_1/V given V is beta(1,n).
$u^*(X) = P[X_1 > 10|V] = P(beta(1, n) > 10/V) = (1 - 10/V)^n$

Eg (conditioning on an insufficient statistic): X_1, X_2 iid N$(\theta, 1)$. Then $\bar{X}$ is unbiased for θ. Let $\phi(\bar(X)) = E(\bar{X}|X_1)$. Then this is unbiased and has lower variance. But it is not an estimator (depends on θ).

4.2 Mean squared Error

Definition 3 *The Mean Squared Error (MSE) of an estimator W of a parameter θ is the function of θ defined by $E_\theta(W - \theta)^2$.*

Alternatively, Mean absolute error or expectation of any other increasing function of $|W - \theta|$ can be used as a measure of performance of an estimator. The advantage of MSE id easy tractability and the interpretation $MSE = Var + Bias^2$. (prove). For an unbiased estimator MSE=var. But a biased estimator might have lower MSE and will be preferred in most cases.

In the iid normal case, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Here $E(S^2) = \sigma^2, var[(n-1)S^2/\sigma^2] = 2(n-1), var(S^2) = 2\sigma^4/(n-1) = mse$. Now let us consider $\hat{\sigma^2}_{MLE} = (n-1)S^2/n$. Bias=$\sigma^2/n$. Var=$2(n-1)\sigma^4/n^2$. MSE=$(2n-1)\sigma^4/n^2$. This is smaller than MSE of the unbiased estimator S^2. Thus by trading off variance for bias, MSE is improved.

Eg Let $X_1, \cdots X_n$ be iid Ber(p). The MLE of p is $\bar{X}$ with MSE=Var=p(1-p)/n.
Consider the Bayes estimator with $Beta(\alpha, \beta)$ prior. The estimator equals $\hat{p}_B(\sum X_i + \alpha)/(n + \alpha + \beta)$. Taking $\alpha = \beta = \sqrt{n}/2$ makes MSE$(\hat{p}_B)$ constant as a function of p. With this prior, for small n, $\bar{X}$ has lower MSE than $\hat{p}_B$ unless p is close to zero or one. For large n, $\hat{p}_B$ has lower MSE than $\bar{X}$ unless p is close to half.

4.3 Information Inequality

Assumptions I. The set $A = x : p(x, \theta) > 0$ does not depend on θ. For all $x \in A, \theta \in \Theta, \partial/partial\theta logp(x, \theta)$ exists and is finite.
II. If T is any statistic such that E($|T|$) ¡ ∞ for all $\theta \in \Theta$, then the operations of integration and differentiation can be interchanged in $\partial/partial\theta \int T(x)p(x, \theta)dx$.

Theorem 4 *If $p(x, \theta) = h(x)exp\{\eta(\theta)T(x) - B(\theta)\}$ is an exponential family and $\eta(\theta)$ has a nonvanishing continuous derivative on Θ, then I and II hold.*

The Fisher Information is defined as $I(\theta) = E(logp(X, \theta))^2$.

Theorem 5 *Suppose that I and II hold and that $E|logp(X, \theta)| < \infty$. Then $E(logp(X, \theta)) = 0$ and $I(\theta) = Var(logp(X, \theta))$.*

Theorem 6 *(Information Inequality/ Cramer Rao Lower Bound) Let T(X) be any statistic such that Var(T(X)) ¡ ∞ for all θ. Denote E(T(X)) by $\psi(\theta)$. Suppose that I and II hold and $0 < I(\theta) < \infty$. Then for all $\theta, \psi(\theta)$ is differentiable and*

$$Var(T(X)) \geq \frac{(\psi'(\theta))^2}{I(\theta)}$$