STAT 450

Lecture 22

Today's notes
Reading for Today's Lecture:
Goals of Today's Lecture:

- Introduce the use of Mean Squared Error to compare estimators.
- Introduce unbiased estimation.
- Derive the Cramér Rao lower bound.

Today's notes

Unbiased Estimation

The problem of the coin and the thumbtack illustrates a general phenomenon. An estimator can be good for some values of θ and bad for others. When comparing $\hat\theta$ and $\tilde\theta$, two estimators of θ we will say that $\hat\theta$ is better than $\tilde\theta$ if it has *uniformly* smaller MSE:

$$MSE_{\hat\theta}(\theta) \le MSE_{\tilde\theta}(\theta)$$

for **all** θ. Normally we also require that the inequality be strict for at least one θ.

The definition raises the question of the existence of a *best* estimate – one which is better than every other estimator. There is no such estimate. Suppose $\hat\theta$ were such a best estimate. Fix a θ^* in Θ and let $\tilde p \equiv \theta^*$. Then the MSE of $\tilde p$ is 0 when $\theta = \theta^*$. Since $\hat\theta$ is better than $\tilde p$ we must have

$$MSE_{\hat\theta}(\theta^*) = 0$$

so that $\hat\theta = \theta^*$ with probability equal to 1. This makes $\hat\theta = \tilde\theta$. If there are actually two different possible values of θ this gives a contradiction; so no such $\hat\theta$ exists.

Principle of Unbiasedness: A good estimate is unbiased, that is,

$$E_\theta(\hat\theta) \equiv 0 \, .$$

WARNING: In my view the Principle of Unbiasedness is a load of hog wash.

For an unbiased estimate the MSE is just the variance.

Definition: An estimator $\hat{\phi}$ of a parameter $\phi = \phi(\theta)$ is **Uniformly Minimum Variance Unbiased** (UMVU) if, whenever $\tilde{\phi}$ is an unbiased estimate of ϕ we have

$$\text{Var}_\theta(\hat{\phi}) \le \text{Var}_\theta(\tilde{\phi})$$

We call $\hat{\phi}$ the UMVUE. ('E' is for Estimator.)

The point of having $\phi(\theta)$ is to study problems like estimating μ when you have two parameters like μ and σ for example.

Cramér Rao Inequality

If $\phi(\theta) = \theta$ we can derive some information from the identity

$$E_\theta(T) \equiv \theta$$

When we worked with the score function we derived some information from the identity

$$\int f(x,\theta)dx \equiv 1$$

by differentiation and we do the same here. If $T = T(X)$ is some function of the data X which is unbiased for θ then

$$E_\theta(T) = \int T(x)f(x,\theta)dx \equiv \theta$$

Differentiate both sides to get

$$\begin{aligned}
1 &= \frac{d}{d\theta}\int T(x)f(x,\theta)dx \\
&= \int T(x)\frac{\partial}{\partial\theta}f(x,\theta)dx \\
&= \int T(x)\frac{\partial}{\partial\theta}\log(f(x,\theta))f(x,\theta)dx \\
&= E_\theta(T(X)U(\theta))
\end{aligned}$$

where U is the score function. Since we already know that the score has mean 0 we see that

$$\text{Cov}_\theta(T(X),U(\theta)) = 1$$

Now remember that correlations are between -1 and 1 or

$$1 = |\mathrm{Cov}_\theta(T(X), U(\theta))| \le \sqrt{\mathrm{Var}_\theta(T)\mathrm{Var}_\theta(U(\theta))}$$

Squaring gives the inequality

$$\mathrm{Var}_\theta(T) \ge \frac{1}{I(\theta)}$$

which is called the Cramér Rao Lower Bound. The inequality is strict unless the correlation is 1 which would require that

$$U(\theta) = A(\theta)T(X) + B(\theta)$$

for some non-random constants A and B (which might depend on θ.) This would prove that

$$\ell(\theta) = A^*(\theta)T(X) + B^*(\theta) + C(X)$$

for some further constants A^* and B^* and finally

$$f(x,\theta) = h(x)e^{A*(\theta)T(x)+B^*(\theta)}$$

for $h = e^C$.

Summary of Implications

- You can recognize a UMVUE sometimes. If $\mathrm{Var}_\theta(T(X)) \equiv 1/I(\theta)$ then $T(X)$ is the UMVUE. For instance in the $N(\mu, 1)$ example the Fisher information is n and $\mathrm{Var}(\overline{X}) = 1/n$ so that $\overline{X}$ is the UMVUE of μ.

- In an asymptotic sense the MLE is nearly optimal: it is nearly unbiased and (approximate) variance nearly $1/I(\theta)$.

- Good estimates are highly correlated with the score function.

- Densities of the exponential form given above are somehow special. (The form is called an exponential family.)

- For most problems the inequality will be strict. It is strict unless The score is an affine function of a statistic T and T (possibility divided by some constant which doesn't depend on θ) is unbiased for θ.

What can we do to find UMVUEs when the CRLB is a strict inequality?

Example: Suppose X has a Binomial(n,p) distribution. The score function is

$$U(p) = \frac{1}{p(1-p)}X - \frac{n}{1-p}$$

Thus the CRLB will be strict unless $T = cX$ for some c. If we are trying to estimate p then choosing $c = n^{-1}$ does give an unbiased estimate $\hat{p} = X/n$ and $T = X/n$ achieves the CRLB so it is UMVU.

A different tactic proceeds as follows. Suppose $T(X)$ is some unbiased function of X. Then we have

$$E_p(T(X) - X/n) \equiv 0$$

because $\hat{p} = X/n$ is also unbiased. If $h(k) = T(k) - k/n$ then

$$E_p(h(X)) = \sum_{k=0}^{n} h(k)\binom{n}{k}p^k(1-p)^{n-k} \equiv 0$$

The left hand side of the $\equiv$ sign is a polynomial function of p as is the right. Thus if the left hand side is expanded out the coefficient of each power p^k is 0. The constant term occurs only in the term $k = 0$ and its coefficient is

$$h(0)\binom{n}{0} = h(0)$$

Thus $h(0) = 0$. Now $p^1 = p$ occurs only in the term $k = 1$ with coefficient $nh(1)$ so $h(1) = 0$. Since the terms with $k = 0$ or 1 are 0 the quantity p^2 occurs only in the term with $k = 2$ with coefficient

$$n(n-1)h(2)/2$$

so $h(2) = 0$. We can continue in this way to see that in fact $h(k) = 0$ for each k and so the *only* unbiased function of X is X/n.

Now a Binomial random variable is just of sum of n iid Bernoulli(p) random variables. If $Y_1, \ldots, Y_n$ are iid Bernoulli(p) then $X = \sum Y_i$ is Binomial(n,p). Could we do better by than $\hat{p} = X/n$ by trying $T(Y_1, \ldots, Y_n)$ for some other function T?

Let's consider the case $n = 2$ so that there are 4 possible values for Y_1, Y_2. If $h(Y_1, Y_2) = T(Y_1, Y_2) - [Y_1 + Y_2]/2$ then again

$$E_p(h(Y_1, Y_2)) \equiv 0$$

and we have

$$E_p(h(Y_1,Y_2)) = h(0,0)(1-p)^2 + [h(1,0)+h(0,1)]p(1-p) + h(1,1)p^2$$

This can be rewritten in the form

$$\sum_{k=0}^{n} w(k)\binom{n}{k}p^k(1-p)^{n-k}$$

where $w(0) = h(0,0)$, $w(1) = [h(1,0)+h(0,1)]/2$ and $w(2) = h(1,1)$. Just as before it follows that $w(0) = w(1) = w(2) = 0$. This argument can be used to prove that for any unbiased estimate $T(Y_1,\ldots,Y_n)$ we have that the average value of $T(y_1,\ldots,y_n)$ over vectors $y_1,\ldots,y_n$ which have exactly k 1s and $n-k$ 0s is k/n. Now let's look at the variance of T:

$$\begin{aligned}
\mathrm{Var(T)} =& E_p([T(Y_1,\ldots,Y_n)-p]^2)\\
=& E_p([T(Y_1,\ldots,Y_n)-X/n+X/n-p]^2)\\
=& E_p([T(Y_1,\ldots,Y_n)-X/n]^2)\\
&+2E_p([T(Y_1,\ldots,Y_n)-X/n][X/n-p])\\
&+E_p([X/n-p]^2)
\end{aligned}$$

I claim that the cross product term is 0 which will prove that the variance of T is the variance of X/n plus a non-negative quantity (which will be positive unless $T(Y_1,\ldots,Y_n)\equiv X/n$). We can compute the cross product term by writing

$$\begin{aligned}
&E_p([T(Y_1,\ldots,Y_n)-X/n][X/n-p])\\
&\qquad=\sum_{y_1,\ldots,y_n}[T(y_1,\ldots,y_n)-\sum y_i/n][\sum y_i/n-p]p^{\sum y_i}(1-p)^{n-\sum y_i}
\end{aligned}$$

We can do the sum by summing over those $y_1,\ldots,y_n$ whose sum is an integer x and then summing over x. We get

$$\begin{aligned}
&E_p([T(Y_1,\ldots,Y_n)-X/n][X/n-p])\\
&=\sum_{x=0}^{n}\sum_{\sum y_i=x}[T(y_1,\ldots,y_n)-\sum y_i/n][\sum y_i/n-p]p^{\sum y_i}(1-p)^{n-\sum y_i}\\
&\qquad=\sum_{x=0}^{n}\left[\sum_{\sum y_i=x}[T(y_1,\ldots,y_n)-x/n]\right][x/n-p]p^x(1-p)^{n-x}
\end{aligned}$$

We have already shown that the sum in [] is 0!

This long, algebraically involved, method of proving that $\hat{p} = X/n$ is the UMVUE of p is one special case of a general tactic.

To get more insight I begin by rewriting

$$
\begin{aligned}
&E_p(T(Y_1,\ldots,Y_n)) \\
&\qquad \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) P(Y_1 = y_1,\ldots,Y_n = y_n) \\
&= \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) P(Y_1 = y_1,\ldots,Y_n = y_n | X = x) P(X = x) \\
&\qquad\qquad = \sum_{x=0}^{n} \frac{\sum_{\sum y_i = x} T(y_1,\ldots,y_n)}{\binom{n}{x}} \binom{n}{x} p^x (1-p)^{n-x}
\end{aligned}
$$

Notice that the large fraction in this formula is the average value of T over values of y when $\sum y_i$ is held fixed at x. Notice that the weights in this average do not depend on p. Notice that this average is actually

$$
\begin{aligned}
&E(T(Y_1,\ldots,Y_n | X = x)) \\
&\qquad \sum_{y_1,\cdots,y_n} T(y_1,\ldots,y_n) P(Y_1 = y_1,\ldots,Y_n = y_n | X = x)
\end{aligned}
$$

Notice that the conditional probabilities do not depend on p. In a sequence of Binomial trials if I tell you that 5 of 17 were heads and the rest tails the actual trial numbers of the 5 Heads are chosen at random from the 17 possibilities; all of the 17 choose 5 possibilities have the same chance and this chance does not depend on p.

Notice that in this problem with data $Y_1,\ldots,Y_n$ the log likelihood is

$$\ell(p) = \sum Y_i \log(p) - (n - \sum Y_i) \log(1-p)$$

and

$$U(p) = \frac{1}{p(1-p)} X - \frac{n}{1-p}$$

as before. Again we see that the CRLB will be a strict inequality except for multiples of X. Since the only unbiased multiple of X is $\hat{p} = X/n$ we see again that $\hat{p}$ is UMVUE for p.

Sufficiency

In the binomial situation the conditional distribution of the data $Y_1, \ldots, Y_n$ given X is the same for all values of θ; we say this conditional distribution is **free** of θ.

Definition: A statistic $T(X)$ is sufficient for the model $\{P_\theta; \theta \in \Theta\}$ if the conditional distribution of the data X given $T = t$ is free of θ.

Intuition: Why do the data tell us about θ? Because different values of θ give different distributions to X. If two different values of θ correspond to the same joint density or cdf for X then we cannot, even in principle, distinguish these two values of θ by examining X. We extend this notion to the following. If two values of θ give the same conditional distribution of X given T then observing T in addition to X does not improve our ability to distinguish the two values.

Mathematically Precise version of this intuition: If $T(X)$ is a sufficient statistic then we can do the following. If $S(X)$ is any estimate or confidence interval or whatever for a given problem but we only know the value of T then:

- Generate an observation X^* (via some sort of Monte Carlo program) from the conditional distribution of X given T.
- Use $S(X^*)$ instead of $S(X)$. Then $S(X^*)$ has the same performance characteristics as $S(X)$ because the distribution of X^* is the same as that of X.

You can carry out the first step **only** if the statistic T is sufficient; otherwise you need to know the true value of θ to generate X^*.

Example 1: IF $Y_1, \ldots, Y_n$ are iid Bernoulli(p) then given $\sum Y_i = y$ the indexes of the y successes have the same chance of being any one of the $\binom{n}{y}$ possible subsets of $\{1, \ldots, n\}$. This chance does not depend on p so $T(Y_1, \ldots, Y_n) = \sum Y_i$ is a sufficient statistic.

example 2: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ then the joint distribution of $X_1, \ldots, X_n, \overline{X}$ is multivariate normal with mean vector whose entries are all μ and variance covariance matrix which can be partitioned as

$$\left[\begin{array}{cc} I_{n\times n} & \mathbf{1}_n/n \\ \mathbf{1}_n^t/n & 1/n \end{array}\right]$$

where $\mathbf{1}_n$ is a column vector of n 1s and $I_{n\times n}$ is an $n \times n$ identity matrix.

You can now compute the conditional means and variances of X_i given $\overline{X}$ and use the fact that the conditional law is multivariate normal to prove that the conditional distribution of the data given $\overline{X} = x$ is multivariate normal with mean vector all of whose entries are x and variance-covariance matrix given by $I_{n\times n} - \mathbf{1}_n\mathbf{1}_n^t/n$. Since this does not depend on μ we find that $\overline{X}$ is sufficient.

WARNING: Whether or not a statistic is sufficient depends on the density function and on Θ.

Rao Blackwell Theorem

Theorem: Suppose that $S(X)$ is a sufficient statistic for some model $\{P_\theta, \theta \in \Theta\}$. If T is an estimate of some parameter $\phi(\theta)$ then:

1. $E(T|S)$ is a statistic.
2. $E(T|S)$ has the same bias as T; if T is unbiased so is $E(T|S)$.
3. $\mathrm{Var}_\theta(E(T|S)) \le \mathrm{Var}_\theta(T)$ and the inequality is strict unless T is a function of S.
4. The MSE of $E(T|S)$ is no more than that of T.

Proof: It will be useful to review conditional distributions a bit more carefully at this point. The abstract definition of conditional expectation is this:

Definition: $E(Y|X)$ is any function of X such that

$$E\left[R(X)E(Y|X)\right] = E\left[R(X)Y\right]$$

for any function $R(X)$.

Definition: $E(Y|X = x)$ is a function $g(x)$ such that

$$g(X) = E(Y|X)$$

Fact: If X, Y has joint density $f_{X,Y}(x, y)$ and conditional density $f(y|x)$ then

$$g(x) = \int y f(y|x) dy$$

satisfies these definitions.

Proof:

$$\begin{aligned}
E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\
&= \int\int R(x)yf(y|x)f_X(x)dydx \\
&= \int\int R(x)yf_{X,Y}(x,y)dydx \\
&= E(R(X)Y)
\end{aligned}$$

You should simply think of $E(Y|X)$ as being what you get when you average Y holding X fixed. It behaves like an ordinary expected value but where functions of X only are like constants.