Lectures 10-17

4 Methods of Point Estimation

[CB7.2, BD2]
Point estimate: Any function of the data. That is, any statistic.
Estimate vs estimator.

4.1 Method of Moments

Definition 1 *Let $X_1, \cdots, X_n$ be iid with pdf (or pmf) $f_\theta, \theta \in \Theta$. We assume that first k moments $m_1, \cdots, m_k$ of f_θ exist. If θ can be written as $\theta = h(m_1, ..., m_k)$, the method of moments estimate of θ is*

$$\hat{\theta}_{MOM} = T(X_1, ..., X_n) = h(\sum_{i=1}^{n} X_i, \cdots, \sum_{i=1}^{n} X_i^k)$$

Note:

- The Definition above can also be used to estimate joint moments. For example, we use $\sum_{i=1}^{n} X_i Y_i$ to estimate $\mathrm{E}(XY)$.
- If θ is not a linear function of the population moments, $\hat{\theta}_{MOM}$ will, in general, not be unbiased. However, it will be consistent and (usually) asymptotically Normal.
- Method of moments estimates do not exist if the related moments do not exist.
- Method of moments estimates may not be unique. If there exist multiple choices for $\hat{\theta}_{MOM}$, one usually takes the estimate involving the lowest–order sample moment.

eg. Normal
eg. Binomial with both n and p unknown.

Example 1 *$X_1, \cdots, X_n$ iid Gamma(p, λ). The first two moments of the gamma distribution are $E(X) = p/\lambda$ and $E(X^2) = p(p+1)/\lambda^2$. Use this to obtain the MOM estimator.*

Example 2 *(Different MoM estimators)Example: $X_1, \cdots, X_n$ iid Poisson(λ). The first moment is λ . Thus, the method of moments estimator based on the first moment is $\bar{X}$. We could also consider using the second moment to form a method of moments estimator.. The method of moments estimator based on the second moment solves $\bar{X^2} = \lambda + \lambda^2$ Solving this equation (by taking the positive root), we find that $\hat{\lambda} = -1/2 + (1/4 + \bar{X^2})^{1/2}$. The two method of moments*

estimators are different. For example, for the data
rpois(10,1) 2 3 0 1 2 1 3 1 2 1,
the method of moments estimator based on the first moment is 1.1 and the method of moments estimator based on the second moment is 1.096872. We choose rthtet one based on the lower moment.

Example 3 *(Hardy-Weinberg proportions) Consider (first generation of) a population in which the alleles A and a are encountered with probabilities θ and 1-θ respectively, $\theta \in (0,1)$. If the alleles are chosen at random and independently for each individual in the next generation, then the probability of having the AA genotype is θ^2, the aa genotype is $(1-\theta)^2$ and Aa genotype $2\theta(1-\theta)$. Suppose we sample n individuals from the population, observe their genotypes and would like to estimate the probability (proportion) of A allele in the population. The corresponding statistical model is an i.i.d. sample $X_1, \cdots, X_n$, where X_i takes values in AA, Aa,AA with probabilities $\theta^2, 2\theta(1-\theta) and (1-\theta)^2$ respectively. Note that $E_\theta N_{AA} = \theta^2$ and $E_\theta N_{aa} = (1-\theta)^2$. Also, $E(N_{AA} + 1/2N_{aA} = \theta$. Each of these can be used to find a method of moments estimator for θ.*

Example 4 *(The method of moments does not use all the information that is available.) $X_1, \cdots, X_n$ iid Uniform(0,θ). The method of moments estimator based on the first moment is $\hat{\theta} = 2\bar{X}$. If $2\bar{X} < max(X_i)$, we know that $\theta > \hat{\theta}$.*

Definition 2 *Suppose we are given a function $\Psi : \Omega \times \mathbb{R}^d \to \mathbb{R}^d$ and define*

$$V(\theta_0, \theta) = E_{\theta_0} \Psi(X, \theta)$$

Suppose $V(\theta_0, \theta) = 0$ has θ_0 as its unique solution for all $\theta_0 \in \Theta$. Then we say $\hat{\theta}$ solving

$$\Xi(X, \hat{\theta}) = 0$$

is an estimating equation estimate.

eg: Take $\Xi = (\hat{\mu}_1 - \mu_1, \cdots, \hat{\mu}_d - \mu_d)$ to get the method of moments estimator.
eg: Least squares as estimating equation.

Definition 3 *Consider a parameter that can be written as a function of F, i.e., $\theta = T(F)$. The plug-in estimator of θ is $T(\hat{F_n}$ where F_n is the empirical cdf.*

For parametric models plug-in estimators are not generally optimal. But they are good starting points for numerical algorithms.
Sample median is a plug-in estimator of population median $theta = F^{-1}(1/2)$, but not a MoM estimator.

4.2 Maximum likelihood estimation

Definition 4 *Let $(X_1, \cdots, X_n)$ be a random vector with pdf (or pmf) $f(x_1, \cdots, x_n; \theta), \theta \in \Theta$. We call the function $L(\theta; x_1, \cdots, x_n) = f(x_1, \cdots, x_n; \theta)$ of θ the likelihood function.*

Definition 5 *A maximum likelihood estimate (MLE) is an estimate $\hat{\theta}_{ML}$ such that*

$$L(\hat{\theta}_{ML}; x_1, \cdots, x_n) = sup_{\theta \in \Theta} L(\theta; x_1, \cdots, x_n).$$

Note: It is often convenient to work with log L when determining the maximum likelihood estimate. Since the log is monotone, the maximum is the same.
Use the derivative to find potential MLE. Use the double derivative to confirm local maximum. Check boundary and confirm global maximum.
If the function is NOT differentiable with respect to θ. Use numerical methods. Or perform directly maximization, using inequalities, or properties of the function.
For multivariate θ, second derivative test for maxima entails checking that the Hessian matrix (matrix of second derivatives) is negative definite. Sign of determinant of any principal minor is $(-1)^r$, where r is the order.
Eg: X_1, X_2, X_3, X_4 i.i.d. Bernoulli(p), $0 < p < 1$. Plot the likelihood function for x = (1, 1, 1, 1), x=(0,0,0,0) and x=(1,1,0,0).
Eg (bivariate parameter) Normal.
Eg (non-unique) Unif($\theta - 1/2, \theta + 1/2$)
Eg Unif(0, θ).
Eg Ber(p), with $\Theta = (1/2, 3/4)$. Here MLE is worse in the sense of MSE to $\hat{p} = 1/2$. Eg: MLE of θ in the Hardy Weinberg set-up.
Eg: MLE(discrete parameter space) Hypergeometric. Total 12, marked θ, pick 5. If $X = 3$ then $\hat{\theta} = 7$.
In this case, MoM estimator does not exist. 12*3/5=7.2 is not in parameter space.
The likelihood function is not a probability mass function or a probability density function: in general, it is not true that L integrates to 1 with respect to θ .
The MLE is the parameter point for which the observed sample is most likely.

Theorem 1 *Let T be a sufficient statistic for $f_\theta, \theta \in \Theta$. If MLE of θ exists, it is a function of T.*

Proof: Since T is sufficient, we can write

$$f(x, \theta) = h(x) g(T(x), \theta)$$

due to the Factorization Criterion. Maximizing the likelihood function with respect to θ takes $h(x)$ as a constant and therefore is equivalent to maximizing $g(T(x), \theta)$ with respect to θ. But $g(T(x), \theta)$ involves x only through T.

- MLE may not exist.
- MLE may not be unique.
- Computation may be difficult.

Theorem 2 *(Invariance of MLE) Let $\{f_\theta : \theta \in \Theta\}$ be a family of pdf's (or pmf's) with $\Theta \subseteq R^k, k \geq 1$. Let $h : \Theta \to \Delta$ be a mapping of Θ onto $\Delta \subseteq, R^p 1 \leq p \leq k$. If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is an MLE of $h(\theta)$.*

Proof: For each $\delta \in \Delta$, we define $\Theta_\delta = \{\theta : \theta \in \Theta, h(\theta) = \delta\}$
and $M(\delta; x) = sup_{\theta \in \Theta_\delta} L(\theta; x)$, the likelihood function induced by h.
Let $\hat{\theta}$ be an MLE let and $\hat{\delta} = h(\hat{\theta})$.
It holds $M(\hat{\delta}; x) = sup_{\theta \in \Theta_{\hat{\delta}}} L(\theta; x) \geq L(\hat{\theta}, x)$ since $\hat{\theta} \in \Theta_{\hat{\delta}}$
But also $M(\hat{\delta}; x) \leq sup_{\delta \in \Delta} M(\delta; x) = sup_{\delta \in \Delta}(sup_{\theta \in \Theta_\delta} L(\theta; x)) = sup_{\theta \in \Theta} L(\theta; x) = L(\hat{\theta}; x)$.
Therefore, $M(\hat{\delta}; x) = L(\hat{\theta}; x) = sup_{\delta \in \Delta} M(\delta; x)$.
Thus, $\hat{\delta} = h(\hat{\theta})$ is an MLE.
eg Let $X_1, ..., X_n$ be iid Ber(p). Let $h(p) = p(1-p)$. Since the MLE of p is $X = \sum(X_i)$, the MLE of h(p) is $X(1-X)$.

4.3 Bayesian methods

Model: $X_, \cdots, X_n \sim f(\mathbf{X}|\theta)$.
In the frequentist approach, θ is a fixed unknown constant.
In the Bayesian approach, we put a prior probability distribution on θ, say $\pi(\theta)$.
The model is then the conditional distribution of the data given a value of θ.
The joint distribution is, therefor the product of the prior and the model.
We use Bayes Rule to obtain the conditional distribution of θ given the data.
This is called the posterior distribution and is given below:

$$f(\theta|\mathbf{X}) = \frac{\text{joint}}{\text{marginalof}\mathbf{X}} = \frac{\pi(\theta) f(\mathbf{X}|\theta)}{\int_{\eta \in \Theta} \pi(\eta) f(\mathbf{X}|\eta) d\eta}.$$

The Bayes estimator is the conditional expectation of θ given the data, that is, the expectation of the posterior distribution and is given by:

$$E(\theta|\mathbf{X}) = \int_{\eta \in \Theta} \eta f(\eta|\mathbf{X}) d\eta = \frac{\int_{\eta \in \Theta} \eta \pi(\eta) f(\mathbf{X}|\eta) d\eta}{\int_{\eta \in \Theta} \pi(\eta) f(\mathbf{X}|\eta) d\eta}.$$

Eg: $X_i \sim iid\mathcal{N}(\theta, 1), \theta \sim \mathcal{N}(0, \sigma^2)$.
Eg Bernoulli with beta(r,r) prior

Definition 6 *A family of prior probability distributions π is said to be conjugate to a family of likelihood functions L(x; θ) if the resulting posterior distributions are in the same family as prior; the prior is called a conjugate prior for the likelihood.*

eg Poisson with Gamma prior as conjugate.

5 Numerical methods for finding MLE's

5.1 Bisection

The bisection method is a method for finding the root of a one-dimensional function that is continuous on $\mathbb{R}$, for which f is monotone increasing or decreasing. It can be used when the likelihood equation is (or can be reduced to)

a one-parameter equation. The bisection method works by repeatedly dividing an interval in half and then selecting the subinterval in which the root exists.

5.2 Coordinate ascent

The coordinate ascent method is an approach to finding the maximum likelihood estimate in a multidimensional family. The coordinate ascent method works by using the bisection method iteratively. Suppose we have a k-dimensional parameter $(\theta_1, \cdots, \theta_k)$. The coordinate ascent method is: Choose an initial estimate $(\hat{\theta}_1, \cdots, \hat{\theta}_k)$.

1. Set $(\hat{\theta}_1, \cdots, \hat{\theta}_k)_{old}$=$(\hat{\theta}_1, \cdots, \hat{\theta}_k)$
2. Maximize $l(\theta_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$ over θ_1 using the bisection method. Reset θ_1 to the value that maximizes the likelihood as $\hat{\theta}_1$.
3. Maximize l over θ_2 using the bisection method. Reset $\hat{\theta}_2$.
4. continue to θ_K
5. Stop if the distance between $(\hat{\theta}_1, \cdots, \hat{\theta}_k)_{old}$ and $(\hat{\theta}_1, \cdots, \hat{\theta}_k)$ is less than some tolerance ϵ. Otherwise return to step 1.

The coordinate ascent method converges to the maximum likelihood estimate when the log likelihood function is strictly concave on the parameter space. See Figure 2.4.1 in Bickel and Doksum.

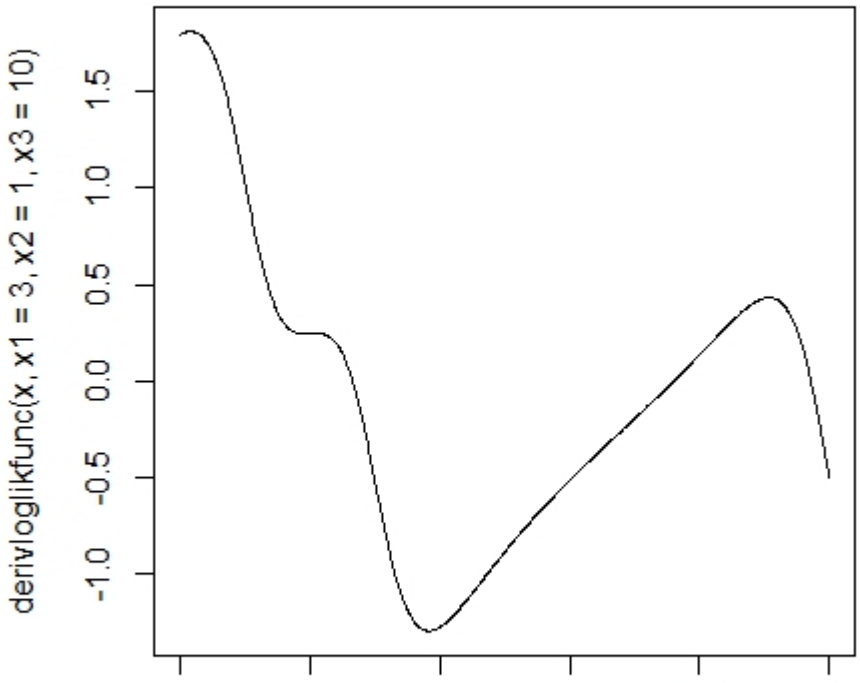
Example (Beta Distribution) This is a two parameter full rank exponential family and hence the log likelihood is strictly concave. We found the method of moments estimates and use them as initial estimates. $\hat{r} = \bar{x}(\bar{x} - \bar{x^2})/(\bar{x^2} - \bar{x}^2)$ $\hat{s} = (1 - \bar{x})(\bar{x} - \bar{x^2})/(\bar{x^2} - \bar{x}^2)$

```
R code for finding the MLE:
# Code for beta distribution MLE
# xvec stores the data
# rhatcurr, shatcurr store current estimates of r and s
# Generate data from Beta(r=2,s=3) distribution)
xvec=rbeta(20,2,3);
#xvec = (0.3184108, 0.3875947, 0.7411803, 0.4044642,  0.7240628, 0.7247060, 0.1091041, 0.138
# Set low and high starting values for the bisection searches
rhatlow=.001;
rhathigh=20;
shatlow=.001;
shathigh=20;
# Use method of moments for starting values
rhatcurr=mean(xvec)*(mean(xvec)-mean(xvec^2))/(mean(xvec^2)-mean(xvec)^2);
shatcurr=((1-mean(xvec))*(mean(xvec)-mean(xvec^2)))/(mean(xvec^2)-mean(xvec)^2);
#rhatcurr=2.239774
#shatcurr=2.893378
```

```
rhatiters=rhatcurr;
shatiters=shatcurr;
derivrfunc=function(r,s,xvec){
  n=length(xvec);
  sum(log(xvec))-n*digamma(r)+n*digamma(r+s);
}
derivsfunc=function(s,r,xvec){
  n=length(xvec);
  sum(log(1-xvec))-n*digamma(s)+n*digamma(r+s);
}
dist=1;
cc=1;
toler=.0001;
while(dist>toler){
  rhatnew=uniroot(derivrfunc,c(rhatlow,rhathigh),s=shatcurr,xvec=xvec)$root;
  shatnew=uniroot(derivsfunc,c(shatlow,shathigh),r=rhatnew,xvec=xvec)$root;
  dist=sqrt((rhatnew-rhatcurr)^2+(shatnew-shatcurr)^2);
  rhatcurr=rhatnew;
  shatcurr=shatnew;
  rhatiters=c(rhatiters,rhatcurr);
  shatiters=c(shatiters,shatcurr);
cc=cc+1}
rhatmle=rhatcurr;
shatmle=shatcurr;
#rhatmle=2.401314
#shatmle=3.117656
#cc=21
```

Example of nonconcave likelihood: Cauchy model. Log likelihood is not concave and has two local maxima between 0 and 10. There is also a local minimum. The local maximum (i.e., the solution to the likelihood equation) that the bisection method finds depends on the interval searched over.

```
R program to use bisection method
derivloglikfunc=function(theta,x1,x2,x3){
  dloglikx1=2*(x1-theta)/(1+(x1-theta)^2);
  dloglikx2=2*(x2-theta)/(1+(x2-theta)^2);
  dloglikx3=2*(x3-theta)/(1+(x3-theta)^2);
  dloglikx1+dloglikx2+dloglikx3;
}
plot(x,derivloglikfunc(x,x1=3,x2=1,x3=10),type="l")
uniroot(derivloglikfunc,interval=c(0,5),x1=3,x2=1,x3=10);
#$root=2.653812
uniroot(derivloglikfunc,interval=c(0,10),x1=3,x2=1,x3=10);
#$root=9.721143
```



5.3 Newton's Method

Newton's method is a numerical method for approximating solutions to equations. The method produces a sequence of values that, under ideal conditions, converges to the MLE . To motivate the method, we expand the derivative of the log likelihood around $\hat{\theta}_{MLE}$: $0 = l'(\hat{\theta}_{MLE}) \approx l'(\theta^{(j)}) + (\hat{\theta}_{MLE} - \theta^{(j)})l''(\theta^{(j)})$ Solving for $\hat{\theta}_{MLE}$ gives $\hat{\theta}_{MLE} = \theta^{(j)} - l'(\theta^{(j)})/l''(\theta^{(j)})$ This suggests the following iterative scheme: $\theta^{(j+1)} = \theta^{(j)} - l'(\theta^{(j)})/l''(\theta^{(j)})$ Newton's method can be extended to more than one dimension (usually called Newton-Raphson) $\theta^{(j+1)} = \theta^{(j)} - \ddot{l}^{-1}(\theta^{(j)})/\dot{l}(\theta^{(j)})$ where $\dot{l}$ denotes the gradient vector of the likelihood and $\ddot{l}$ denotes the Hessian.

Comments on methods for finding the MLE:

1. The bisection method is guaranteed to converge if there is a unique root in the interval being searched over but is slower than Newton's method.
2. Newton's method does not work if $l''(\theta^{(j)}) \approx 0$
3. Newton's method does not always converge.
4. For the coordinate ascent method and Newton's method, a good choice of starting values is often the method of moments estimator or plug-in estimator.

5. When there are multiple roots to the likelihood equation, the solution found by the bisection method, the coordinate ascent method and Newton's method depends on the starting value. These algorithms might converge to a local maximum (or a saddlepoint) rather than a global maximum.

5.4 EM Algorithm

Complete data, incomplete data.
E step: Expectation of complete data log likelihood given incomplete data.
M step: Maximize
Iterate.
This is a famous example from Rao (1973)[Linear Statistical Inference and Its Applications]. We consider the genetic linkage of 197 animals, in which the phenotypes are distributed into 4 categories: Y = (y1, y2, y3, y4) = (125, 18, 20, 34) with cell probabilities $(1/2 + \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$.
Though it is by no means impossible to maximize this multinomial likelihood directly, we illustrate how the EM algorithm brings a substantial simplification, by using the augmentation method. Specifically, we augment the observed data Y by dividing the first cell into two, with respective cell probabilities 1/2 and $\theta/4$. This gives an augmented data set X = (x1, x2, x3, x4, x5), where x1 + x2 = y1, and x3 = y2, x4 = y3, x5 = y4.
E-step: $E(l) = (E(X2) + x5)log(\theta) + (x3 + x4)log(1-\theta)$.
$X_2 \sim Bin(y1, \theta/(\theta+2))$
M-step: $\theta_{n+1} = (159\theta_n + 68)/(197\theta_n + 144)$
The alternation between estimation and maximization is clearly seen in this iteration formula. Starting with $\theta_0 = 0.5$ we obtain the sequence as follows 0.6082, 0.6243, 0.6265, 0.6268, 0.6268.