Maximum Likelihood Estimation

Neba Nfonsang

University of Denver

Introduction to MLE

- The principle of maximum likelihood is attributed to the work of R.A Fisher(1950).
- Suppose we have N independently identically distributed(iid) random variables X=[X1, X2, …., XN] and a corresponding vector or sample of observed data, $x = [x_1, x_2, x_3, \cdots, x_n]$, the likelihood of the parameter θ is the joint probability density of the observed data under different parameter settings $\theta = [\theta_1, \theta_2, \ldots, \theta_p]$.

Eliason (193)

The probability density function

- Note that a function that maps a value or a range of values to their probability of occurrence is known as probability function for discrete variables or probability density function for continuous variables.
- However note that statisticians call the probability function for both discrete and continuous random variables as probability density function.
- Sometimes the probability function for discrete variables is called probability mass function.

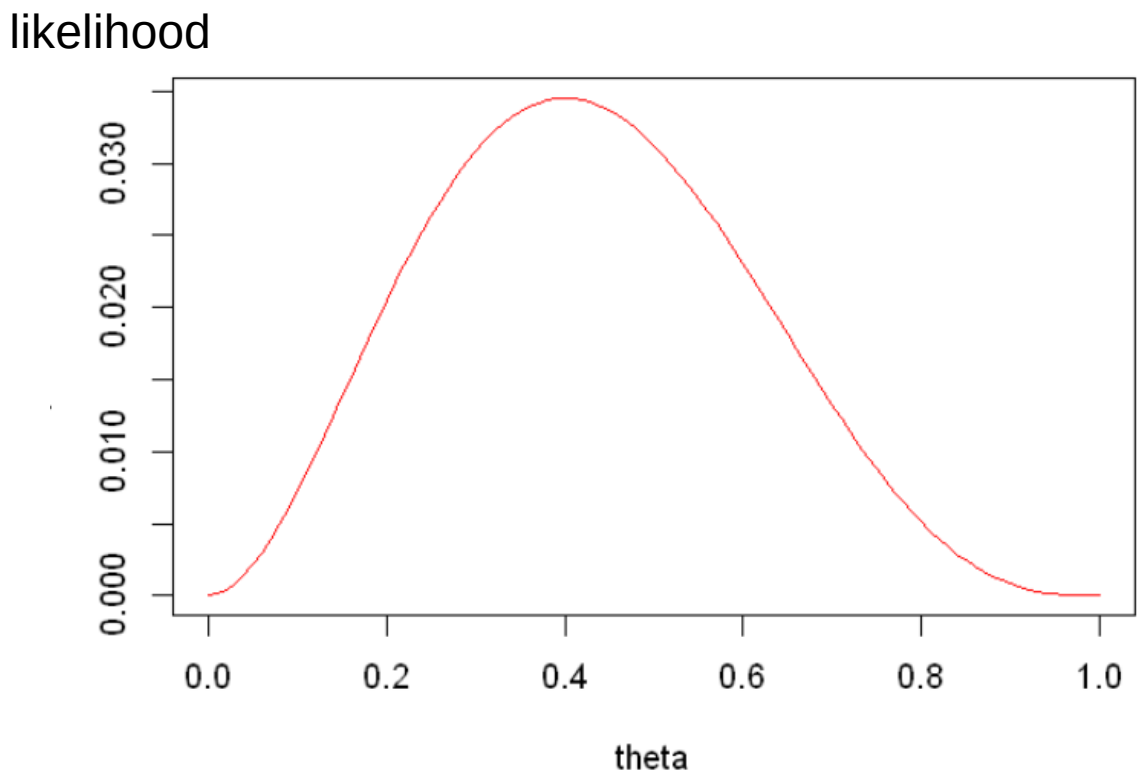
Likelihood

- "*the likelihood L(θ) is the probability of the observed data x considered as a function of θ." (*Pawitan, 2001).
- The likelihood of the parameter, $l(\theta)$, is the joint probability density of the observed data as a function of the parameter.
- The generic data x could be a set or sample of observations from an experiment of any complexity, a vector of values, a matrix, an array of matrices, a time series, a range of values instead exact measurements (Pawitan, 2001).

Likelihood

Likelihood function is the joint probability of the data as a function of the parameter.

- For a single parameter, and a single variable, the likelihood function could be as shown on the graph.
- If there are 2 successes in 5 trials in a binomial distribution, then likelihood function would look like this. The theta value that maximizes the likelihood function is 0.4.



MLE

- Given an iid sample of observations $x = [x_1, x_2, x_3, \cdots, x_n]$ from a probability distribution f(x; θ), the value of θ that maximizes the likelihood function given the data $l(\theta/X)$ is the maximum likelihood estimate.

- So, we want to find an estimate for θ that maximizes the likelihood of observing the data.
- The likelihood of a parameter is the joint probability of observing the data given the parameters.

MLE

- *The likelihood function is the joint probability density of the data as a function of the parameter.*
- *MLE(θ) is the parameter estimate that maximizes the likelihood function given the data*
- $l(\theta|X) = f(X;\theta)$
- $\boldsymbol{l(\theta|X) = \prod_{i=1}^{n} f(x_i;\theta)}$

$$= f\ (x_1;\theta)^{*} f(x_2;\theta) \ldots f(x_n;\theta)$$

MLE

- **Log-Likelihood Function:**
- Log $l(\theta|X) = \log(\prod_{i=1}^{n} f(x_i|\theta))$

$=\log(f(x_1|\theta)*f(x_2|\theta)\dots f(x_n|\theta))$

$= \log f(x_1|\theta) + \log f(x_2|\theta) \dots \log f(x_n|\theta)$

- $\boldsymbol{\log l(\theta|X) = \sum_{i=1}^{n} log\, f(x_i|\theta)}$
- $\text{MLE}(\theta) = \text{maximize}\, l(\theta|X) = \text{maximize}\, \log(l(\theta|X))$
  - Find $\frac{\partial \log l(\hat{\theta}|x)}{\partial \theta} = 0$

MLE Example: parameter of binomial

- Given a single observation x (number of successes) from a binomial distribution with know parameter n and unknown parameter θ, the probability density is defined by:
- $\mathrm{P}(X = x) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{(n-x)}$
- $l(\theta) = \mathrm{P}(x) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{(n-x)}$
- $\log l\,(\theta) = \log(\frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{(n-x)})$

$The\ paramter\ \theta$ is the probability of success in a single trial and n is the total number of trials.

MLE Example: parameter of binomial

- $\log l\,(\theta) = \log(\frac{n!}{x!(n-x)!}\theta^{x}(1-\theta)^{(n-x)})$
- $= \log(\frac{n!}{x!(n-x)!}) + x log\theta + (n-x)\log(1-\theta)$
- $\mathrm{MLE}(\theta)$: $\frac{\partial \log l(\hat{\theta}|x)}{\partial\theta} = 0 + \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$

$$x(1-\theta) = \theta(n-x)$$

$$x \cancel{- x\theta} = \theta n \cancel{- \theta x}$$

$$\boldsymbol{\hat{\theta} = \frac{x}{n}}$$

MLE Example: parameter of Poisson

- Supposed the samples or observations [5, 9, 3] are drawn from a Poison distribution, with a probability density defined as:
- $\mathrm{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- where x is number of events per unit time
- The parameter λ is the average number of events occurring per unit time.
- $\mathrm{P}(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

MLE Example: parameter of Poisson

- $l(\lambda/\mathrm{x}) = \Pi_{1=1}^{n} \mathrm{P}(x;\lambda) = \Pi_{1=1}^{n} \frac{\lambda^{x} e^{-\lambda}}{x!}$ and x= [5, 9, 3]

- $l(\lambda/\mathrm{x}) = \frac{\lambda^{5} e^{-\lambda}}{5!} * \frac{\lambda^{9} e^{-\lambda}}{9!} * \frac{\lambda^{3} e^{-\lambda}}{3!}$

- $l(\lambda/\mathrm{x}) = \frac{\lambda^{5} * \lambda^{9} * \lambda^{3} e^{-\lambda-\lambda-\lambda}}{5!*9!*3!} = \frac{\lambda^{5+9+3} e^{-3\lambda}}{5!*9!*3!} = \frac{\lambda^{17} e^{-3\lambda}}{5!*9!*3!}$

MLE Example: parameter of Poisson

- $l(\lambda/\mathrm{x}) = \frac{\lambda^{17} e^{-3\lambda}}{5!*9!*3!}$
- $\log l\,(\lambda) = \log(\frac{\lambda^{17} e^{-3\lambda}}{5!*9!*3!}) = \log \lambda^{17}$ - 3λ - log $5! * 9! * 3!$
- $\mathrm{MLE}(\hat{\lambda})$: $\frac{\partial \log l(\lambda|x)}{\partial \theta} = \frac{17}{\lambda}$ - 3 – 0 = 0

$\hat{\lambda} = \frac{17}{3}$ = 5.67 (average number of events per unit time)

MLE Example: parameters of Normal Dist.

- Let $x = [x_1, x_2, x_3, \cdots, x_n]$ be an iid sample from $N(\mu, \delta^2)$.
- The contribution of x_i, to the likelihood is:

$$L_i(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\}$$

- The likelihood function is:
- $l(\theta) = \Pi_{i=}^{n} \frac{1}{\delta\sqrt{2\pi}} \exp\left(-\frac{x_i-\mu}{2\delta^2}\right)^2$
- $l(\theta) = (\frac{1}{2\pi\delta^2})^{n/2} \exp \sum_i^n \left(-\frac{x_i-\mu}{2\delta^2}\right)^2$
- $\log l\,(\theta) = \sum_{i=1}^{n} \log\, l_i(\theta)$
- $\log l\,(\theta) = \log(\frac{1}{2\pi\delta^2})^{n/2} - \sum_i^n \left(-\frac{x_i-\mu}{2\delta^2}\right)^2$

MLE Example: parameters of Normal Dist.

- $l(\theta) = \Pi_{i=}^{n} \frac{1}{\delta\sqrt{2\pi}} \exp\left(-\frac{x_i-\mu}{2\delta^2}\right)^2$
- $l(\theta) = (\frac{1}{2\pi\delta^2})^{n/2} \exp\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\delta^2}\right]$
- $\log l\,(\theta) = \sum_{i=1}^{n} \log\, l_i(\theta) = \frac{n}{2}\log(\frac{1}{2\pi\delta^2}) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\delta^2}$

$$= -\frac{n}{2}\log(2\pi\delta^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\delta^2}$$

MLE Example: parameters of Normal Dist.

- $\log l\,(\mu) = -\frac{n}{2}\log(2\pi\delta^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\delta^2}$
- $\frac{\partial \log l(\mu)}{\partial \mu} = 0 + 2 * \frac{\sum_{i=1}^{n}(x_i-\mu)}{2\delta^2}$

$$\frac{\sum_{i=1}^{n}(x_i-\hat{\mu})}{\delta^2} = 0$$

$$\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\hat{\mu} = 0$$

$$\sum_{i=1}^{n} x_i - n\hat{\mu} = 0 \rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Finding the parameter (μ)

Note that:

$\theta = [\mu, \delta^2]$

MLE Example: parameters of Normal Dist

- $\log l\,(\delta) = -\frac{n}{2}\log(2\pi\delta^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\delta^2}$
- $\frac{\partial \log l(\delta)}{\partial \delta} = -\frac{n}{2} * 2 * \frac{2\pi\delta}{2\pi\delta^2} + 2 * \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\delta^3}$

$$= -\frac{n}{\hat{\delta}} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\widehat{\delta^3}} = 0$$

$$n = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\widehat{\delta^2}}$$

$$\widehat{\delta^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \longrightarrow \boldsymbol{\hat{\delta} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}}$$

Continuous Model

- For continuous variables, observing x is similar to observing $x \in (x - {}^{\epsilon}/_{2},\ x + {}^{\epsilon}/_{2})$. That means, x has some measurement error, so x is called the precision limit.
- If ϵ is small enough, on observing x the likelihood of θ is:

$$\begin{aligned} L(\theta) &= P_\theta\{X \in (x - \epsilon/2, x + \epsilon/2)\} \\ &= \int_{x-\epsilon/2}^{x+\epsilon/2} p_\theta(x)dx \approx \epsilon\ p_\theta(x). \end{aligned}$$

ϵ can be ignored if we are comparing θ within the model

"Hence, *in all continuous models where the outcome x is observed with good precision we will simply use the density function pθ (x) to compute the likelihood."*

(Pawitan, 2001).

Assumptions

- We have just seen that MLE assumes small measurement errors so, if the data is have huge measurement errors, inaccurate and invalid data, then we violate the assumption for MLE for continuous data.
- This is why it is important to clean the data and prepare it for analysis so we can have valid results.
- Also, the data must be iid. This is a simplifying assumption that makes it possible to use the joint probability density for a sample of observations.

Discrete Model

- For a discrete model, the probability density function is used directly and this represents the probability function of the parameter as well.

Using Optimize in R

- For a single random variable, the optimize function in R can be easily used to estimate the parameter that maximizes the likelihood function.

- Given that 2 successes in 5 trials follow a binomial distribution
- Estimate the parameter of the model that maximizes the likelihood function

```
likelihood <- function(theta) {
    like <- (theta^2*(1-theta)^3)
}
# note that we did not include the 5C3 because it is a constant
# and when we compare theta within the same model, we can drop
# that term to simplify the math.
```

```
optimize(likelihood, interval = c(0,1), maximum = T)
```

```
$maximum
0.399999427550837
$objective
0.034559999999882
```

```
# similar to
2/5
```

Poisson Example in R

Example 2

Supposed the samples or observations [5, 9, 3] are drawn from a Poison distribution. Find the parameter estimate that is consistent with this data

$$\mathrm{P}(X = x) = \frac{\Lambda^x e^{-\Lambda}}{x!}$$

- a poisson distribution models the probability of the number of events per unit time, x, and the parameter is the average number of events per unit time.

Example in R

```
pois.likelihood <- function(lambda){
    x <- c(5, 9, 3)
    like <- 1
    for (i in x){
       like <- like*((lambda^i)*exp(-lambda))/factorial(i)
    }

   return (like)
}
```

```
optimize(pois.likelihood, interval = c(3, 9), maximum = T)
# supply the minimum and maximum value of the data in this case
# for the interval since the average should lie in this range.
```

$maximum
5.66666447249428
$objective
0.001015006004751O1

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

```
# alternative calculation
x <- c(5, 9, 3)
sum(x)/length(x)
```

5.66666666666667

Multiple Parameters in R

Model with Multiple Parameters

When there are multiple parameters, the nlm() function can be used to optimize the likelihood function. Since nlm() performs minimization, it would rather be good to minimize the negative log likelihood which is the same as maximizing the likelihood function.

Problem Frame

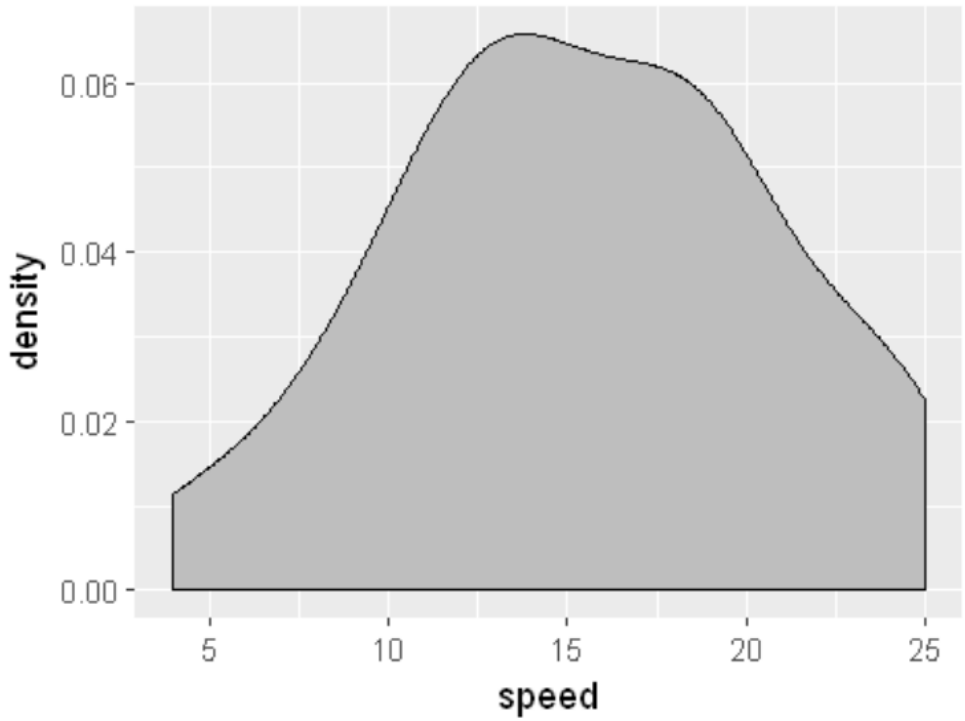
Let's assume that the values of speed from the cars data are an iid sample and follow a normal distribution with mean mu and standard deviation signma. Using Maximum Likelihood Estimation, estimate the value of the parameters mu and sigma that maximize the likelihood function.

```
x <- cars$speed
x
```

```
 4  4  7  7  8  9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15 15 16 16 17 17 17 18 18 18 18
19 19 19 20 20 20 20 20 22 23 24 24 24 24 25
```

Data is assumed to be an iid sample

```
require(tidyverse, quiet=T)
options(repr.plot.width=4, repr.plot.height=3)
ggplot(data=cars, aes(x=speed)) +
stat_density(fill="gray", color="black")
```



Estimating Multiple Parameters

```r
# define the negative log likelihood function
neg.loglike <- function(thetha, sample){
    return (-sum(log(dnorm(sample, mean=thetha[1], sd=thetha[2]))))
}


# minimize the negative log likelihood
#supply the function, starting parameters
# avoid current state (starting parameters) being too close to zero
nlm(neg.loglike, c(20, 20), sample=x)
```

$minimum
153.710501995084

$estimate
15.4000000122119 5.2345009697827

$gradient
2.22852039675313e-08 -1.88410189037745e-08

$code
1

$iterations
13

```r
# actual mean and standard deviation
mean(x)
sd(x)
```

15.4

5.28764443523478

The Assumption of Likelihood

- MLE assumes the data is iid. This is a simplifying assumption to the joint probability of the data.
- Joint probability of the data given that the data is dependent can be written as a chain of conditional probability also know as the chain rule

$$P(A_1, A_2, A_3 \ldots, A_n) = P(A_1 | A_2, A_3, .., A_n) P(A_2 | A_3, \ldots, A_n) \ldots P(A_n)$$

OR

$$P(A_1, \ldots, A_n) = P(A_1) P(A_2 | A_1) P(A_3 | A_1, A_2) \ldots P(A_n | A_1, \ldots, A_{n-1})$$

$$= \prod_{i=1}^{n} P(A_i | A_1, \ldots, A_{i-1})$$

Assuming that the sample data is iid, the joint probability is simplified to the product of the densities of the data.