Jan Izquierdo
Eloi Vilella

Maximum Likelihood Practice

ML estimation of a one-parameter distribution.

1. (16p) ML estimation of a one-parameter distribution. Let X be a random variable with probability density f(x | β) = βe−βx+1 with x > 1/ β We consider a random sample of n observations of this distribution.

a) (2p) Write down the likelihood function for a sample of n observations of this distribution

$$L(\beta|x) = \prod_{i=1}^{n} \beta \cdot e^{-\beta x + 1}$$

b) (1p) Obtain the log-likelihood function.

We solve using logarithmic properties:

$$l(\beta \mid x) = \log(L(\beta \mid x)) = \log\left(\prod_{i=1}^{n} \beta e^{-\beta x + 1}\right)$$

$$\log(L(\beta|x)) = n\log(\beta) - \beta\sum_{i=1}^{n} x + n$$

c) (2p) Find the stationary point(s) of the log-likelihood function analytically

$$\frac{\partial}{\partial \beta} l(\beta \mid x) = \frac{n}{\beta} - \sum_{i=1}^{n} x$$

$$\widehat{\beta} = \frac{n}{\sum_{i=1}^{n} x} \rightarrow \theta = [0, 1]$$

$$\frac{n}{\sum_{i=1}^{n} x} \rightarrow \textit{stationary point in } \theta = [0, 1]$$

d) (1p) Determine whether the stationary point(s) are maxima or minima

Jan Izquierdo
Eloi Vilella

$$\frac{\partial 2}{\partial \beta} \log(L(\beta \mid x)) = -\frac{n}{\beta} < 0$$

Since the second derivative is negative, the stationary point is a maxima.

e) (1p) Download the file Sample.dat, which contains sample of observations from this probability distribution. Determine the sample size and calculate the value of the ML estimator for this sample. (See Sample.dat file)

Refer to R script output for more information

$wc -w =100000

$wc -l =100000

After importing the data to Rstudio the number of observations is 100000. 100000 is sample size.

$$\frac{d(ln(\beta))}{d(\beta)} = 0.47$$ <- MLE soluciona això

f) (2p) Plot the log-likelihood function, and assess graphically if your ML estimate coincides with the maximum of this function.

Refer to R script output for more information

g) (1p) Determine an expression for the Fisher information by calculating

$$I_x(\beta) = -E(\frac{\partial^2 l}{\partial \beta^2}) = -E[\frac{\partial^2 l}{\partial \beta^2} ln(L(\beta|x))] = -E[-\sum_{i=1} x_i]$$

Probability of X depends on B, L(B|x) is the probability density function of X where X is conditioned by B.
This other method yields the same results.

$$I_x(\beta) = -E(\frac{\partial^2 l}{\partial \beta^2}) = -E[\frac{\partial^2}{\partial \beta^2} log(L(\beta))] = -E[-\sum_{i=1} x_i]$$

I(β) is fisher information,which is the expected value of the second derivative with respect to β.

Jan Izquierdo
Eloi Vilella

h) (1p) Use the Fisher information for obtaining an expression for the variance of the maximum likelihood estimator βˆML.

$$Var(\hat{\beta}) \geq \frac{1}{I_x(\beta)}$$

$$Var(\hat{\beta}) = \frac{1}{\sum_i x_i}$$

i) (1p) Using the asymptotic normality of the ML estimator, give an expression of a confidence interval for β.

$$C1_{1-a} = \hat{\beta} \pm z_{a/2}\sqrt{V(\hat{\beta})}$$

$$C1_{1-a} = \hat{\beta} \pm z_{a/2}\frac{1}{\sqrt{\sum_i x_i}}$$

j) (1p) Calculate a 95% confidence interval for parameter β, using the dataset that you have downloaded.

Output of the R script:
Confidence interval of 95% for parameter B is [0.4708948 , 0.4764564]

k) (1p) Do you think it is tenable that β = 1? Justify your answer.

β = 1 is not feasible since this value falls outside the confidence interval we've just calculated, which is [0.4708948, 0.4764564]. For β to be valid, it should lie within this interval, considering the results from our previous exercise.

l) 2p) Make a histogram of the data, using function hist, using the argument freq=FALSE. Overplot the histogram with the estimated probability density f(x|β), using the maximum likelihood estimate. What do you observe?

Refer to R script output for more information

2.The Hardy-Weinberg law in genetics says that the proportions of genotypes AA, Aa, and aa are o^2, 2o(1-o), and (1-o)^2, respectively, where o is between 0 and 1. Suppose that in a sample of n individuals from the population (small relative to the

Jan Izquierdo
Eloi Vilella

size of the population), we observe X_1 individuals of type AA, X_2 individuals of type Aa, and X_3 individuals of type aa.

a)What distribution do the counts (X_1, X_2, X_3) follow?

A multinomial distribution as they describe the expected frequencies of different genotypes in the same population.

b) Record the likelihood function, the log-likelihood function, and the score function for o.

n= sample size

Likelihood

$$L(o) = \left(\frac{n!}{x_1!*x_2!*x_m!}\right)\prod_{i=1}^{m} P(x_i)^{x_i} = \frac{n!}{(X_1!*X_2!*X_3!)} * P(X_1)^{X_1} * P(X_2)^{X_2} * P(X_3)^{X_3}$$

$$L(o) = \frac{n!}{(X_1!*X_2!*X_3!)} * o^{2X_1} * (2o(1-o))^{X_2} * (1-o)^{2X_3}$$

$$L(o) = n!\prod_{i=1}^{m}\frac{P(x_i)^{x_i}}{x_i!} = n!\left(\frac{P(X_1)^{X_1}*P(X_2)^{X_2}*P(X_3)^{X_3}}{X_1!*X_2!*X_3!}\right)$$

$$L(o) = n!\left(\frac{o^{2X_1}*(2o(1-o))^{X_2}*(1-o)^{2X_3}}{X_1!*X_2!*X_3!}\right)$$

Log likelihood

$$log(L(o)) = log\left(n!\left(\prod_{i=1}^{m}\frac{P(x_i)^{x_i}}{x_i!}\right)\right) = log\left(n!\left(\frac{P(X_1)^{X_1}*P(X_2)^{X_2}*P(X_3)^{X_3}}{X_1!*X_2!*X_3!}\right)\right)$$

$$L(o) = log\left(n!\left(\frac{o^{2X_1}*(2o(1-o))^{X_2}*(1-o)^{2X_3}}{X_1!*X_2!*X_3!}\right)\right) =$$

$$= log(n!) + (X_1 log(o^2) + X_2 log(2o(1-o)) + X_3 log(1-o)^2) -$$

$$- (log(X_1)! + log(X_2)! + log(X_3)!) = log(L(o))$$

Score function

$$S(o) = \frac{d}{do} log(L(o))$$

c) Record the form of the MLE for o.

Refer to R script output for more information

3. (1p) Find an example of a published research where they study/apply/use MLE, include the link, with the retrieved data and do a summary of 5-6 text lines.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0259111

Jan Izquierdo
Eloi Vilella

Summary:

The study discussed a method for estimating probability density using various flexible basis functions like polynomials, trigonometric functions, and splines. They used the BIC to pick the right type and number of basis functions. The method outperformed other density estimators in various tests, showing better accuracy, modality detection, and computation time. It can also be extended to multivariate cases using suitable basis functions and cross-variable dependencies.

R script Assignment 1

Maximum Likelyhood Estimation

Jan Izquierdo, Eloi Vilella

2023-10-20

Exercise 1

Setting libraries and some variables

```
library("ggplot2")
db<-read.table("./Assignment_data/A1_MLE.dat")
m<-mean(db$V1)
sd1<-sd(db$V1)
variance<-var(db$V1)
```

***1.e)* Download the file Sample.dat, which contains sample of observations from this probability distribution.Determine the sample size and calculate the value of the ML estimator for this sample.** .

```
n <- length(db$V1)
x <- db$V1
b <- seq(min(db$V1),max(db$V1),by=0.01)
L <- (b^n)*(prod(x)^(b-1))
cat("The sample size is", n)
```

```
## The sample size is 100000
```

This method does not work as the sample size (n) is too big, it approaches infinity. We will calculate it this other way.

```
#n/sum(x) x is all values
MLE<-100000/sum(db$V1)
#MLE=~sd1/m
cat("MLE is", MLE)
```

```
## MLE is 0.4736756
```

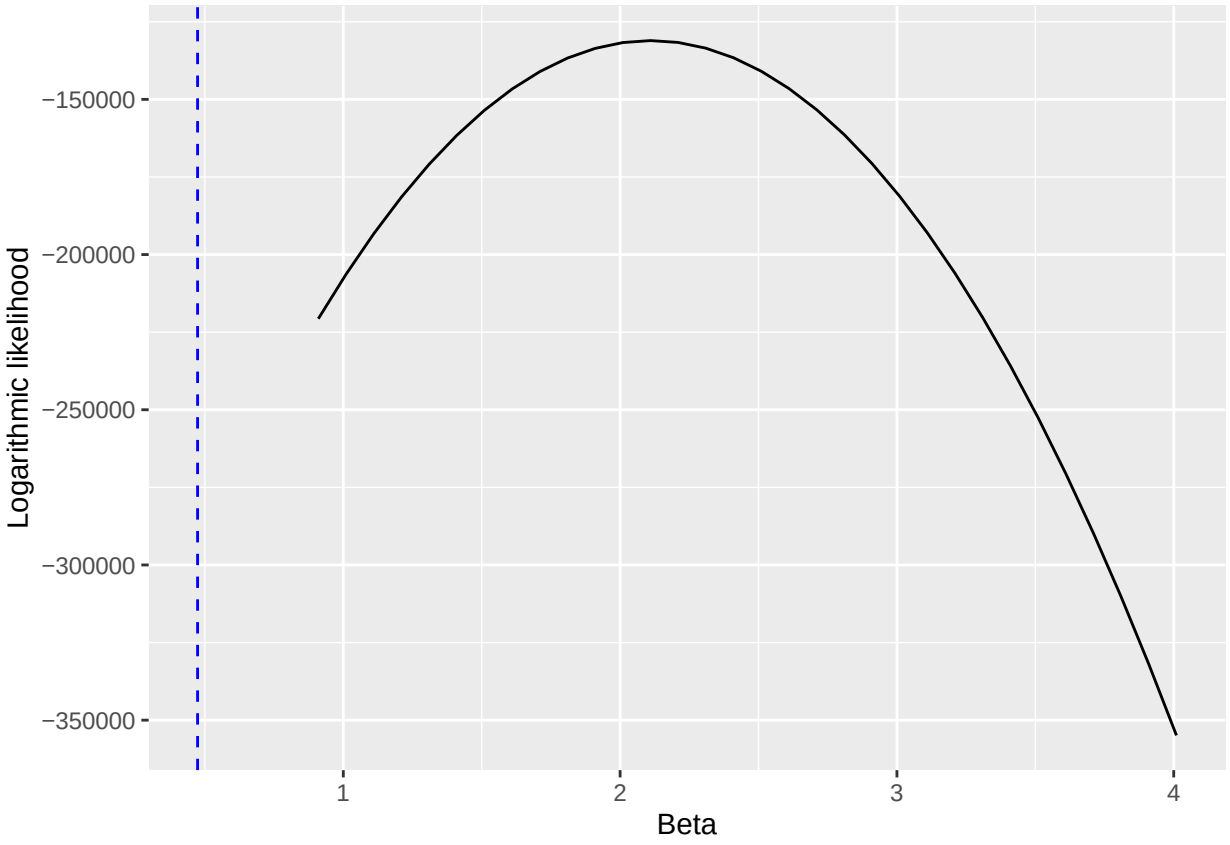
***1.f)* Plot the log-likelihood function, and assess graphically if your ML estimate coincides with the maximum of this function.** .

```
log_likelihood<-function(m){
  log_data<-dnorm(db$V1, mean=m, sd=sd1, log=T)
  log_likelyhood<-sum(log_data)
  return(log_likelyhood)
}

beta<-seq(min(db$V1), max(db$V1), by=0.1)#min->max or 0.1->max
log_likelihood_values <- sapply(beta, log_likelihood)
```

```r
plot_data <- data.frame(Beta = beta, LogLikelihood = log_likelihood_values)

ggplot(plot_data, aes(x=beta, y=log_likelihood_values))+
          geom_line()+
          geom_vline(xintercept =100000/sum(db$V1), color = "blue", linetype = "dashed")+
          xlab("Beta")+
          ylab("Logarithmic likelihood")
```



The blue line represents where the MLE stands respect the graph

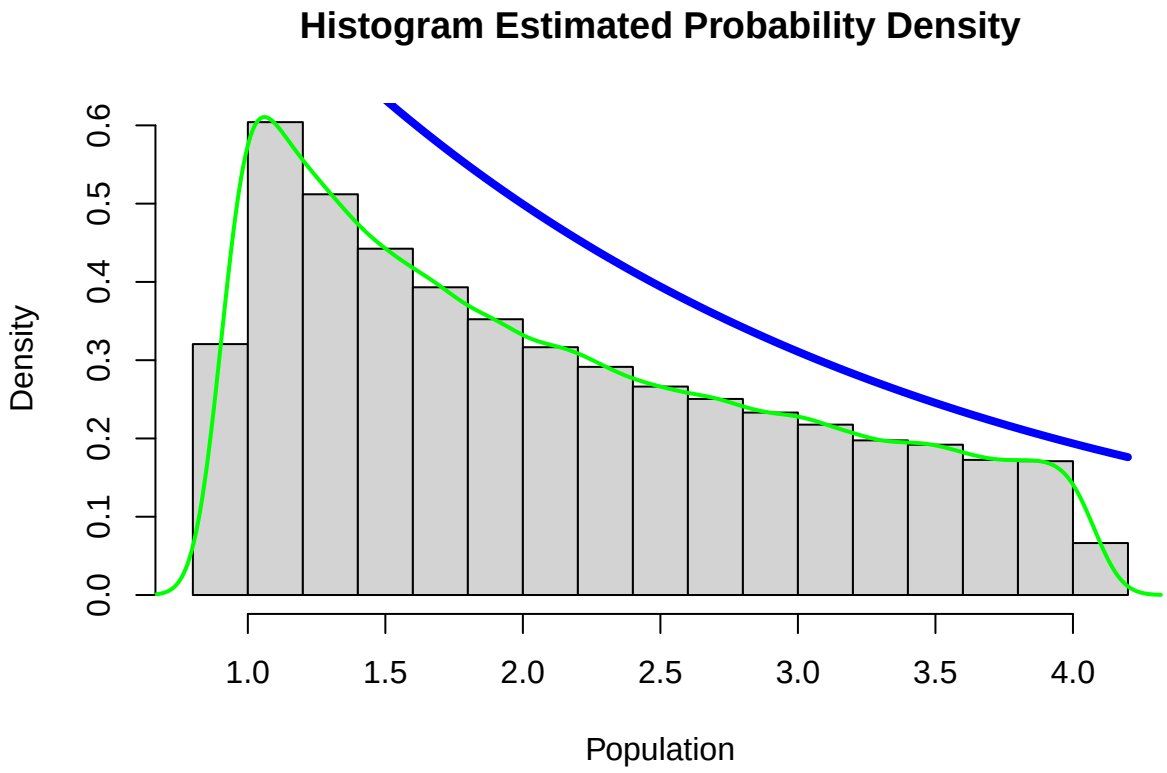
```r
#n is sample size, MLE is maximum likelihood estimator
error <- qt(0.975, df=n-1)*sd1/(2*sqrt(n))
lower_bound <- MLE - error
upper_bound <- MLE + error
cat("Confidence interval of 95% for parameter B is","[", lower_bound, ",", upper_bound, "]")
```

1.j) **Calculate a 95% confidence interval for parameter B, using the dataset that you have downloaded.**

```
## Confidence interval of 95% for parameter B is [ 0.4708948 , 0.4764564 ]
```

1.l) **Make a histogram of the data, using function hist, using the argument freq=FALSE. Overplot the histogram with the estimated probability density f(x|B), using the maximum likelihood estimate. What do you observe?.** .

```r
Prob_density<- function(a, beta) {
  beta * exp(-beta * a + 1)
}
hist(db$V1, freq=F, main = "Histogram Estimated Probability Density", xlab="Population")
lines(density(db$V1), col="green", lwd=2)
curve(Prob_density(x, MLE), col = "blue", lwd = 4, add = TRUE)
```



Exercise 2

Setting libraries and variables

```r
o<-1#placeholder
AA<-o^2
Aa<-2*o*(1-o)
aa<-(1-o)^2
#n<-sample size
#z<-individuals AA
#x<-individuals Aa
#y<-individuals aa
```

***2.c)* Record the form of MLE for o.** .

```r
log_likelihood<-function(o, z ,x ,y){
  n<-z+x+y
  AA<-o^2
  Aa<-2*o*(1-o)
```

```r
  aa<-(1-o)^2
  log_lik<-log(factorial(n))-(log(factorial(z))+log(factorial(x))+log(factorial(y)))+
    z*log(AA)+x*log(Aa)+y*log(aa)
  #z*log(AA)=total individuals* prob of an individual=individuals in sample
  return(-log_lik)#negative, as optimize gets minimum, we want maximum, so we use the negative of the f
}
z<-20#example data
x<-50#example data
y<-10#example data
mle_opt<-optimize(log_likelihood, interval =c(0,1), z, x, y)
mle_o<-mle_opt$minimum
cat("MLE for o is", mle_o)
```

```
## MLE for o is 0.5625006
```