Maximum Likelihood Estimation and Model Fitting, CMEE MSc. Tin-Yu Hui

Practical 3 (24 Feb 2021)

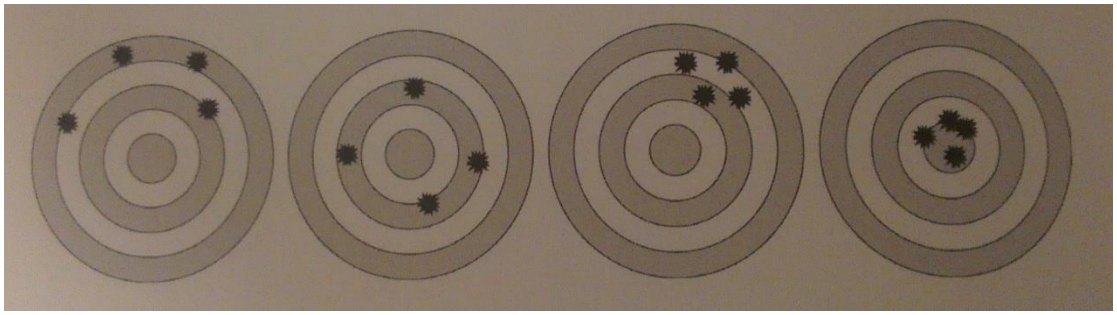
Question 1

What are the properties of maximum likelihood estimators?

Question 2

Accuracy: Measure of bias

Precision: Measure of variability ("spreadness")



[photo credit: The Signal and the Noise: Why so many predictions fail-but some don't]

Relate the graphs to the statements below:

- Accurate but not precise
- Precise but not accurate
- Not accurate and precise
- Both accurate and precise

Question 3

Coin tossing example revisited. If we observe y heads from n independent tosses, then the likelihood function for the unknown parameter p is $L(p) = \binom{n}{y} p^y (1-p)^{n-y}$.

i. If we observed $y = 7$ from $n = 10$ tosses, please perform a likelihood ratio test (by hand) to test for $H_0: p = 0.5$ vs $H_1: p \neq 0.5$ at 5% significance level.

ii. Instead if we observed $y = 35$ from $n = 50$ tosses, the MLE is still the same ($\hat{p} = \frac{7}{10} = \frac{35}{50} = 0.7$). Please perform a likelihood ratio test to test for $H_0: p = 0.5$ vs $H_1: p \neq 0.5$ at 5% significance level.

iii. (To be discussed tomorrow) Find the 95% confidence interval for p for case (ii) above using R.

Question 4

A logistic regression example adopted from Mick Crawley's GLM course. Here we would like to investigate whether the survival of a plant can be determined from the number of flowers and the size of root. The hypothesis (from literature review? theory? common sense? or simply Mick's speculation?) is that plants with more flowers are more likely to die in the following winter, and plants with bigger roots are less likely to die.

In this dataset `flowering.txt` the response variable is `State`, which indicates the current status for a particular plot of plant (`1=alive`, `0=dead`). `Flowers` and `Root` are the two continuous explanatory variables as described above. First let us load the dataset into R and call it `flowering`:

```
flowering<-read.table('flowering.txt', header=T)
names(flowering)
```

Then we make some simply plots to visualise the dataset:

```
par(mfrow=c(1,2))
plot(flowering$Flowers, flowering$State)
plot(flowering$Root, flowering$State)
```

Because all the responses are just 0's and 1's we get two rows of points: one across the top of the plot at `y=1` (alive) and another across the bottom at `y=0` (dead). The plots of binary data are not very informative; fitting a logistic regression seems to more appropriate in this case. The first part of this exercise is to fit a simple logistic regression with the given dataset. The formulae for the logistic regression model are given below (or page 6-9 from Wednesday's slides):

$$y_i \sim Bernoulli(p_i), \text{ where}$$

$$p_i = expit(a + b * Flowers_i + c * Root_i)$$

Now let us construct the log-likelihood for the model above, with the following name `logistic.log.likelihood`

```
# TWO ARGUMENTS: parm IS A VECTOR OF PARAMETERS,
# dat IS THE INPUT DATASET
logistic.log.likelihood<-function(parm, dat)
{
# DEFINE PARAMETERS
a<-parm[1]
b<-parm[2]
c<-parm[3]

# DEFINE RESPONSE VARIABLE, WHICH IS THE FIRST COLUMN OF dat
State<-dat[,1]
```

```
# SIMILARLY DEFINE OUR EXPLANATORY VARIABLES
Flowers<-dat[,2]
Root<-dat[,3]

# MODEL OUR SUCCESS PROBABILITY, VIA EXPIT TRANSFORMATION
p<-exp(a+b*Flowers+c*Root)/(1+exp(a+b*Flowers+c*Root))

# THE LOG-LIKELIHOOD FUNCTION
log.like<-sum(State*log(p)+(1-State)*log(1-p))

return(log.like)
}
```

We may wish to try whether our log-likelihood function has been defined properly. This can be done by evaluating it at an arbitrarily set of parameter values (e.g. `c(0,0,0)`):

```
# TRY
logistic.log.likelihood(c(0,0,0), dat=flowering)
```

It should return a number of around -40.89. This is the log-likelihood value evaluated at `a=0`, `b=0` and `c=0` given our observed data. Next we would like to maximise our log-likelihood function with `optim()`. Create an object `M1` to store the output.

```
# MAXIMISE THE LOG-LIKELIHOOD
M1<-optim(????????????????????????????????????)
M1
```

`M1` is a `list` containing multiple elements. Use the `$` sign to retrieve the elements, say `M1$par` and `M1$value`. What are the MLE for the parameters? What is the associated log-likelihood value?

Some further suggested that the interaction between `Flowers` and `Root` may also play a role in determining `State`. In this slightly more complex model, p_i becomes:

$$p_i = expit(a + b * Flowers_i + c * Root_i + d * Flowers_i * Root_i)$$

$Flowers_i * Root_i$ is the interaction term with its regression coefficient d. This model needs another log-likelihood function:

```
logistic.log.likelihood.int<-function(parm, dat)
{
```

```
# DEFINE PARAMETERS, ONE MORE THIS TIME
??????????????????????????????????????????????

# DEFINE RESPONSE VARIABLE, WHICH IS THE FIRST COLUMN OF dat
State<-dat[,1]

# DEFINE EXPLANATORY VARIABLES
??????????????????????????????????????????????

# MODEL OUR SUCCESS PROBABILITY
p<-??????????????????????????????????????????????

# THE LOG-LIKELIHOOD FUNCTION FOR A SINGLE DATA POINT
# WHICH IS EXACTLY THE SAME
log.like<-sum(State*log(p)+(1-State)*log(1-p))

# THE OVERALL LOG-LIKELIHOOD IS THE SUM OF THE LOG-LIKELIHOODS OF
THE OBSERVATIONS
return(log.like)
}
```

Maximise the log-likelihood function and store the outputs as `M2`.

Using `M1` and `M2`, can you perform a likelihood-ratio test to test for the interaction term at $\alpha = 5\%$ significance level?