Maximum likelihood estimation

HaiYing Wang

Outline

1. A simple example: Coin tosses
2. Likelihood
3. Maximum likelihood
4. Log-likelihood
5. Shape of the log-likelihood
6. Logistic regression
7. Newton's Method
8. Property of the MLE

- Coin tosses have binary outcomes: a Head (H) or a Tail (T).
- Assume that different coin tosses don't impact each other.
- In statistics, this implies that coin toss outcomes are independent and identically distributed, or i.i.d..
- Assume that the coin is biased.
- Let the probability of getting a Head be p and the probability of getting a Tail be $1-p$.
- So how do we find that value of p?

Let's toss the coin five times, and assume that we get the following sequence: H,T,T,H,H.
The probability of seeing this result is

$$L(p) = p(1-p)(1-p)pp = p^3(1-p)^2$$

where 3 is the number of Heads and 2 is the number of Tails.
More generally, if we have a total of N tosses, out of which n are Heads, then the probability is written in a generic function form:

$$L(p) = p^n(1-p)^{N-n}$$

- Here, $L(p)$ is the likelihood of observing the data.
- It is a function of the unknown parameter p.
- We want to use the maximizer of $L(p)$ to estimate p.
- The maximum likelihood estimator (MLE) is

$$\hat{p} = \arg\max_p L(p).$$

The estimation problem now becomes an optimization problem.

- Here we can differentiate L with respect to p and set it equal to zero to find the optimal value of p

$$L'(p) = \frac{dL}{dp} = 0$$

- This will give us a complicated expression:

$$L'(p) = np^{n-1}(1-p)^{N-n} - (N-n)(1-p)^{N-n-1}p^n = 0$$

- This equation is not easy to solve, neither analytically nor numerically.

Now we consider the log-likelihood function.

- Let $l(p) = \log\{L(p)\}$, namely,

$$l(p) = n \log p + (N - n) \log(1 - p).$$

- The maximizer of $L(P)$ is the same as the maximizer of $l(p)$.
- Thus $\hat{p}$ is the solution to

$$l'(p) = \frac{\partial l}{\partial p} = \frac{n}{p} - \frac{N - n}{1 - p} = 0,$$

which is

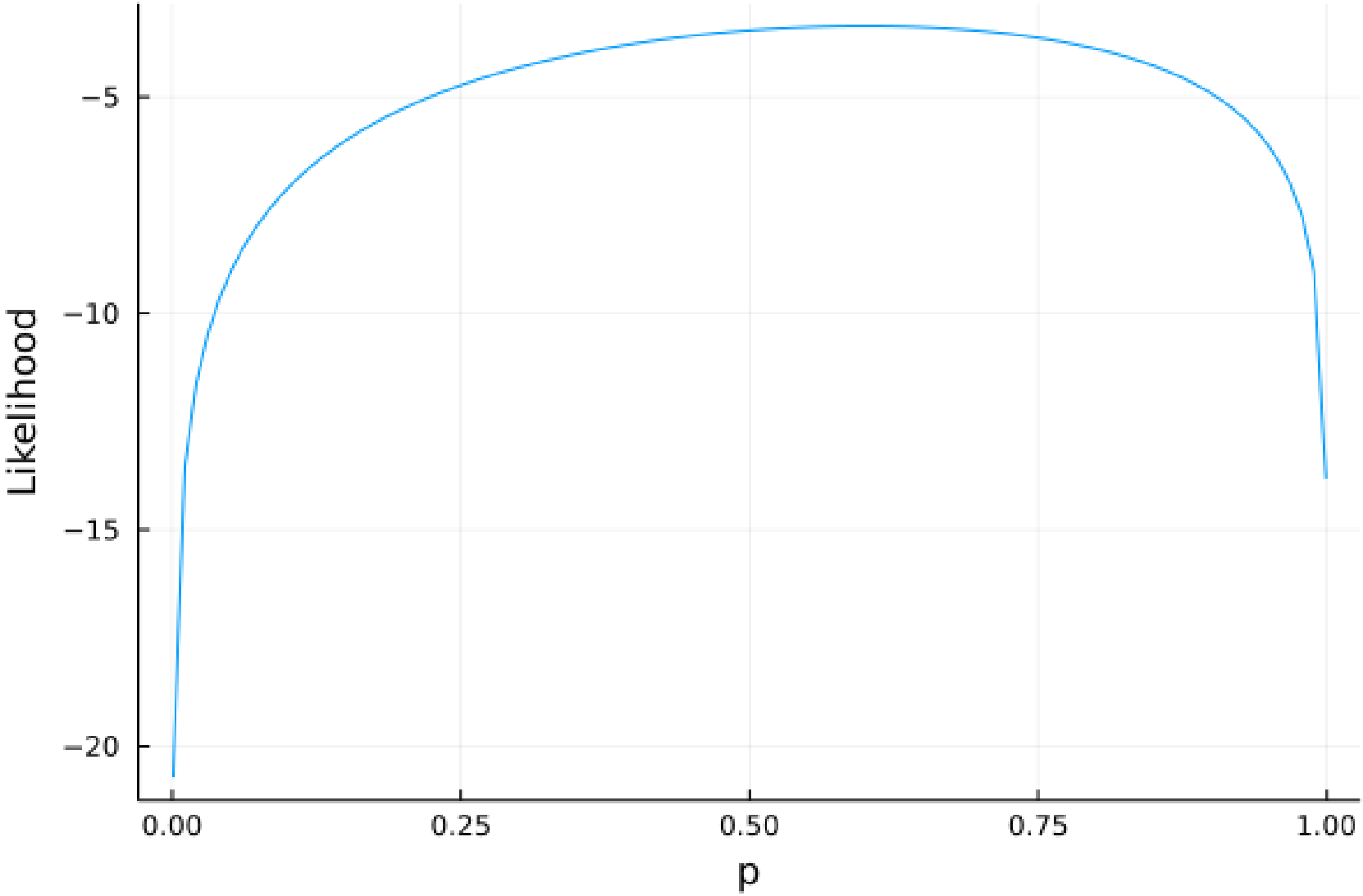
$$\hat{p} = \frac{n}{N} = \frac{3}{5} = 0.6.$$

How to quantify the uncertainty of this $\hat{p}$?

Let's plot the log-likelihood function.

```
1  n = 3
2  N = 5
```

```
p̂ = n/N
l(p) = n * log(p) + (N-n) * log(1-p)
ps = collect(LinRange(0.001, 0.999, 100))
using Plots
plot(ps, l.(ps), xlab="p", ylab="Likelihood", legend=false)
savefig("likelihood.png")
```


Likelihood
−5
−10
−15
−20
0.00
0.25
0.50
0.75
1.00
p

Solving the problem numerically

```
using Optim
l(p) = -3 * log(p) - 2 * log(1 -p)
res = optimize(l, 0, 1)
summary(res)
Optim.minimum(res)
Optim.minimizer(res)
```

```
0.6000000004335264
```

Let $y \in \{0, 1\}$ be binary, and given covariate x, the probability for $y = 1$ is

$$p_x(\theta) = \Pr(y = 1 \mid x) = \frac{e^{x^T\theta}}{1 + e^{x^T\theta}}. \tag{1}$$

For a given data $(x_i, y_i), i = 1, ...N$, the log-likelihood for θ is

$$l(\theta) = \sum_{i=1}^{N} \{y_i \log p_{x_i} + (1 - y_i) \log(1 - p_{x_i})\} \tag{2}$$

$$= \sum_{i=1}^{N} \{y_i x_i^T \theta + \log(1 + e^{x_i^T \theta})\}. \tag{3}$$

The MLE is $\hat{\theta} = \arg\max_p l(\theta)$. How should we find it? Solving

$$l'(\theta) = \sum_{i=1}^{N} \{y_i - p_{x_i}(\theta)\} x_i = 0. \tag{4}$$

Example: Income

An income data set was extracted from the 1994 Census database. There are totally $48,842$ observations in this data set, and the response variable is whether a person's income exceeds \$50K a year.
Can we classify if a person's income exceeds \$50K if we know the following covariates:

- x_1, age;
- x_2, final weight (Fnlwgt);
- x_3, highest level of education in numerical form;
- x_4, capital loss (LosCap);
- x_5, hours worked per week.

A fast approach of root find for a differentiable function, say $g(x) = 0$. The methods starts from some initial value x_0, and for $t = 0, 1, ...,$

compute

$$x_{t+1} = x_t - \left\{ \frac{\partial g(x_t)}{\partial x_t^T} \right\}^{-1} g(x_t)$$

until x_t converges.
The method is based on a linear expansion of $g(x)$. The method is also known as Newton–Raphson iteration. It needs an initial value x_0. If $g(x) = 0$ has multiple solutions, the end result depends on x_0.
Applied to optimization of l, this method solves $g = l' = 0$ and it requires the Hessian l'', which can be difficult to obtain, especially for multivariate functions. Many variants of Newton's method avoid the computation of the Hessian.
For example, to obtain MLE for logistic regression with likelihood $l(\theta)$,

$$l'(\theta) = \sum_{i=1}^{N} \{y_i - p_{x_i}(\theta)\} x_i; \tag{5}$$

$$l''(\theta) = -\sum_{i=1}^{N} w_i(\theta^{(t)}) x_i x_i^T. \tag{6}$$

where $w_i(\theta) = p_i(\theta)\{1 - p_i(\theta)\}$.
Thus we can obtain the MLE by iteratively applying the following

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \left\{l''(\theta^{(t)})\right\}^{-1} l'(\theta^{(t)}). \tag{1}$$

The above iterative formula is not restricted to logistic regression.
How good is the MLE?
Consider a random sample $X_1, \ldots, X_N$ of size N coming from a distribution with density function $f(x|\theta)$, where θ is a parameter vector. The MLE $\hat{\theta}_N$ of the unknown parameter θ is obtained by

maximizing the loglikelihood function

$$l(\theta) = \sum_{i=1}^{N} l_i(\theta) = \sum_{i=1}^{N} \log f(X_i|\theta)$$

with respect to θ. Typically, $\hat{\theta}_N$ is the solution to the score equation $l'(\theta) = 0$.
Let $l''(\theta)$ be the Hessian. Under very mild assumptions

$$\hat{\theta}_N \overset{a}{\sim} N\left(\theta, \{-l''(\theta)\}^{-1}\right). \tag{7}$$

The MLE is the most efficient estimator among all unbiased estimators; its asymptotic variance is the smallest among all unbiased estimators. The asymptotic variance of the MLE can be estimated by inverting the observed Fisher information matrix $-l''(\hat{\theta}_N)$, i.e., $\{-l''(\hat{\theta})\}^{-1}$. This is available at the convergence of the Newton's method.

M-estimator

- More generally in Statistics, M-estimators are a broad class of extremum estimators obtained by maximizing or minimizing an data dependent objective function.

- Both non-linear least squares and maximum likelihood estimation are special cases of M-estimators.
- The definition of M-estimators was motivated by robust statistics, which contributed new types of M-estimators.
- When the objective function is smooth, the M-estimator can be obtained by solving the corresponding "score" equation.
- Clearly, optimization or root-finding are very important in Statistics.