Introduction to maximum likelihood estimation

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

24 April 2020

Goals for today

- Understand the concept of a likelihood function
- Understand the difference between probability and likelihood
- Understand maximum likelihood estimation
- Understand the characteristics of maximum likelihood estimates

Maximum likelihood estimation (MLE)

What is maximum likelihood estimation?

A method used to estimate the parameter(s) of a model given some data

As the name suggests, the goal is to *maximize* the likelihood

The likelihood function

Here we are referring to the likelihood of some parameters given some data, which can be written as

$$\mathcal{L}(\theta|y) \;\text{ or }\; \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})$$

The likelihood function

Here we are referring to the likelihood of some parameters given some data, which can be written as

$$\mathcal{L}(\theta|y) \text{ or } \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})$$

We'll write this as

$$\mathcal{L}(y;\theta) \text{ or } \mathcal{L}(\mathbf{y};\boldsymbol{\theta})$$

to avoid confusion with the "|" meaning *conditional* probability

The likelihood function

Let's define the likelihood function to be

$$\mathcal{L}(y;\theta) = f_\theta(y)$$

where $f_\theta(y)$ is a model for y with parameter(s) θ

The likelihood function

For *discrete* data, $f_\theta(y)$ is the *probability mass function* (pmf)

The likelihood function

For *discrete* data, $f_\theta(y)$ is the *probability mass function* (pmf)

For *continuous* data, $f_\theta(y)$ is the *probability density function* (pdf)

The pmf of pdf can be for *any* distribution

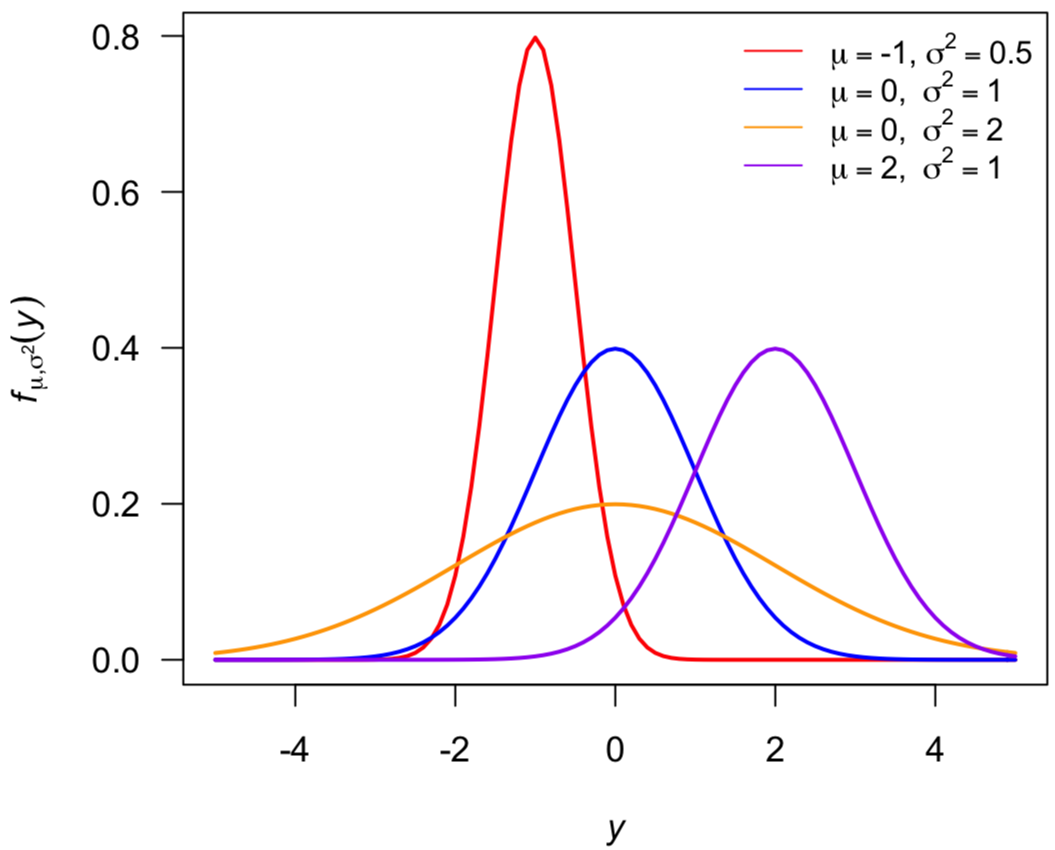
Gaussian likelihood function

Let's begin with the pdf for a Gaussian (normal) distribution

$$f(y; \mu, \sigma^2) \sim \mathrm{N}(\mu, \sigma^2)$$

$$f(y; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[- \frac{(y - \mu)^2}{2\sigma^2} \right]$$

Gaussian likelihood function



Gaussian likelihood function

Note that $f(y; \mu, \sigma^2)$ is ***not*** a probability!

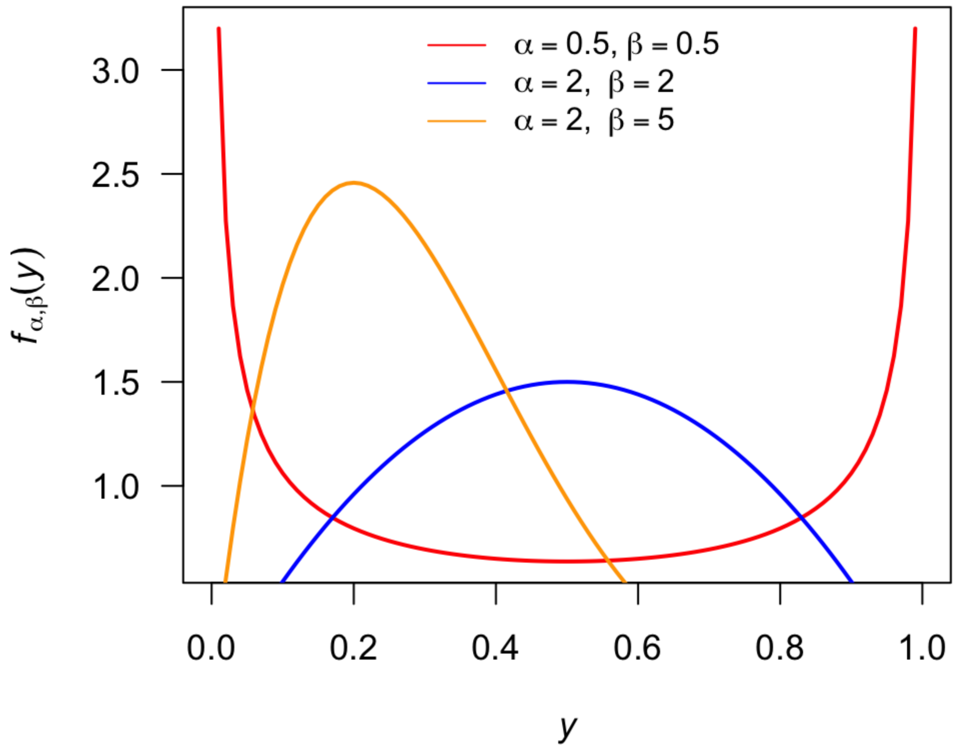
The pdf gives you *densities* for given values of y, μ & σ^2

It's only constraint is

$$\int_{-\infty}^{+\infty} f(y)dy = 1$$

Beta likelihood function

For example, many densities of $\mathrm{Beta}(\alpha, \beta) > 1$



Likelihood vs probability

Probability is linked to *possible results*

Possible results are mutually exclusive and exhaustive

Likelihood vs probability

Probability is linked to *possible results*

Likelihood is linked to *hypotheses*

Hypotheses are neither mutually exclusive nor exhaustive

Likelihood vs probability

An example

- Suppose I ask you to predict the outcomes of 10 tosses of a fair coin
- There are 11 *possible results* (0 to 10 correct predictions)
- The *actual result* will always be only 1 of 11 possible results
- Thus, the probabilities for each of the 11 possible results must sum to 1

Likelihood vs probability

An example

- Suppose you predict 7 of 10 tosses correctly
- I might hypothesize that you just guessed, but someone else might hypothesize that you are a psychic
- These are different hypotheses, but they are not mutually exclusive (you might be a psychic who likes to guess)
- We would say that my hypothesis is nested within the other

Likelihood vs probability

An example

- Importantly, there is no limit to the hypotheses we (or others) might generate
- Because we don't generally consider the entire suite of *all* possible hypotheses, the likelihoods of our hypotheses do not have any absolute meaning
- Only the *relative likelihoods* ("likelihood ratios") have meaning

Maximizing the likelihood

What does it mean to maximize $\mathcal{L}(y; \theta)$?

We want to find the parameter(s) θ of our model $f_\theta(y)$ which are most likely to have generated our observed data y

Maximizing the likelihood

More formally, we can write this as

$$\begin{aligned}\hat{\theta} &= \max_{\theta} \mathcal{L}(y; \theta)) \\ &= \max_{\theta} f_{\theta}(y)\end{aligned}$$

Maximizing the likelihood

In practice, we have multiple observations $y = \{y_1, y_2, \ldots, y_n\}$, so we need the *joint distribution* for y

$$\hat{\theta} = \max_{\theta} f_\theta(y_1, y_2, \ldots, y_n)$$

Maximizing the likelihood

Remember *independent and identically distributed* (IID) errors?

If the data Y are independent, we can make use of

$$f_\theta(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} f_\theta(y_i)$$

The joint probability of all of the y_i is the product of their marginal probabilities

Maximizing the likelihood

If the data Y are identically distributed, we can use the same distribution and parameterization for $f_\theta(y)$

Maximizing the likelihood

If the data Y are both independent and identically distributed, then we have

$$\hat{\theta} = \max_{\theta} \prod_{i=1}^{n} f_{\theta}(y_i)$$

(This assumption isn't necessary, but it makes our lives easier)

Maximum likelihood estimates

The value(s) of $\hat{\theta}$ that maximizes the likelihood function is/are called the *maximum likelihod estimate(s)* (MLE) of θ

Binomial distribution

Let’s begin with a simple example of coin tossing

Assume we have a “fair” coin with equal chance of coming up heads or tails

$\Pr(H) = \Pr(T)$

Binomial distribution

If we flip the coin 2 times, what is the probability that we get exactly 1 heads?

Our 4 possible outcomes are

1. $\{H, H\}$
2. $\{H, T\}$
3. $\{T, H\}$
4. $\{T, T\}$

2 of 4 flips are heads, so $\Pr(H = 1) = 2/4 = 0.5$

Binomial distribution

Let’s think about this in terms of the probabilities

1. $\{H, H\}$: $\Pr(H) \times \Pr(H) = 0.5 \times 0.5 = 0.25$ X
2. $\{H, T\}$: $\Pr(H) \times \Pr(T) = 0.5 \times 0.5 = 0.25$ ✓
3. $\{T, H\}$: $\Pr(T) \times \Pr(H) = 0.5 \times 0.5 = 0.25$ ✓
4. $\{T, T\}$: $\Pr(T) \times \Pr(T) = 0.5 \times 0.5 = 0.25$ X

$\Pr(H = 1) = 0.25 + 0.25 = 0.5$

Binomial distribution

We can generalize this by

1. $\{H, H\} \,:\, \Pr(H) \times \Pr(H)$
2. $\{H, T\} \,:\, \Pr(H) \times (1 - \Pr(H))$
3. $\{T, H\} \,:\, (1 - \Pr(H)) \times \Pr(H)$
4. $\{T, T\} \,:\, (1 - \Pr(H)) \times (1 - \Pr(H)))$

$$\begin{aligned} \Pr(H = 1) &= \Pr(H)(1 - \Pr(H)) + (1 - \Pr(H)) \Pr(H) \\ &= 2[\Pr(H)(1 - \Pr(H))] \end{aligned}$$

Binomial distribution

Now consider the probability of exactly 1 heads in 3 coin tosses

$\{H, H, H\}$ X

$\{H, H, T\}$ X

$\{H, T, H\}$ X

$\{T, H, H\}$ X

$\{H, T, T\}$ ✓

$\{T, H, T\}$ ✓

$\{T, T, H\}$ ✓

$\{T, T, T\}$ X

$$\begin{aligned}
\Pr(H = 1) &= \Pr(H)(1 - \Pr(H))(1 - \Pr(H)) \\
&\quad + (1 - \Pr(H)) \Pr(H)(1 - \Pr(H)) \\
&\quad + (1 - \Pr(H))(1 - \Pr(H)) \Pr(H) \\
&= 3[\Pr(H)(1 - \Pr(H))^2]
\end{aligned}$$

Binomial distribution

Let's define k to be the number of "successes" out of n "trials" and p to be the probability of a success

We can generalize our probability statement to be

$$\Pr(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial distribution

What is the probability of getting 1 heads in 3 tosses?

```
## trials
n <- 3
## successes
k <- 1
## probability of success
p <- 0.5
## Pr(k = 1)
choose(n, k) * p^k * (1 - p)^(n-k)
```

```
## [1] 0.375
```

Binomial distribution

What is the probability of getting 1 heads in 3 tosses?

```
## trials
n <- 3
## successes
k <- 1
## probability of success
p <- 0.5
## Pr(k = 1)
dbinom(k, n, p)
```

```
## [1] 0.375
```

Binomial likelihood

What if we don't know what p is?

For example, we tag 100 juvenile fish in June and 20 are alive the following year

What is the probability of surviving?

Binomial likelihood

We need to find p that maximizes the likelihood

$$\mathcal{L}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\Downarrow$$

$$\max_p \mathcal{L}(20; 100, p) = \binom{100}{20} p^{20} (1-p)^{100-20}$$

Binomial likelihood

Let’s try some different values for p

$$\mathcal{L}(20; 100, 0.3) = \binom{100}{20} 0.3^{20}(1-0.3)^{100-20} \approx 0.0076$$

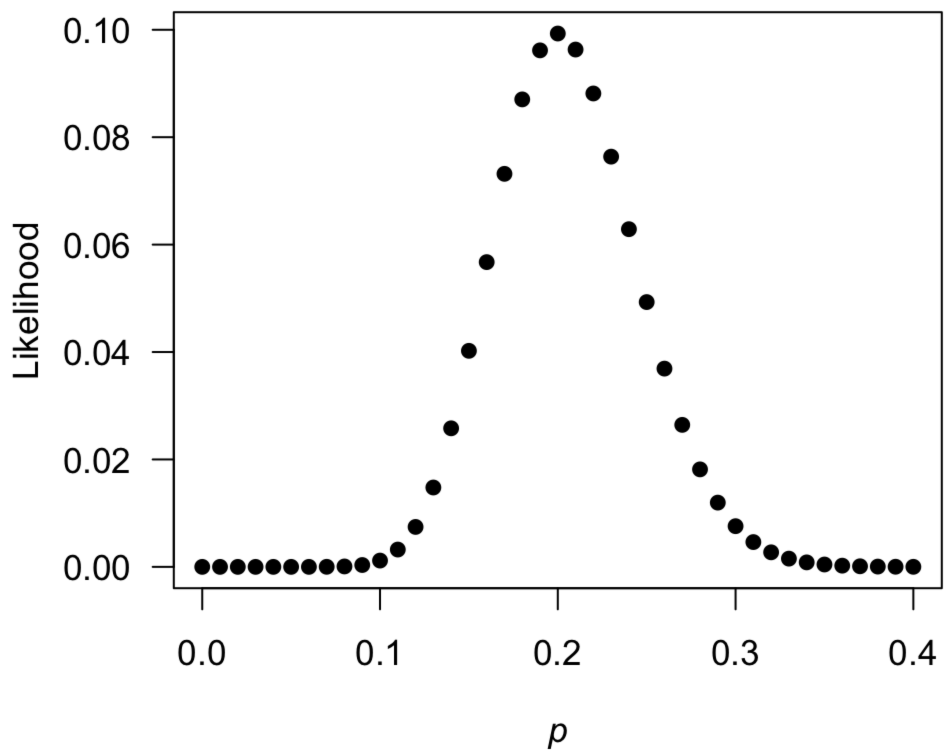
$$\mathcal{L}(20; 100, 0.25) = \binom{100}{20} 0.25^{20}(1-0.25)^{100-20} \approx 0.049$$

$$\mathcal{L}(20; 100, 0.2) = \binom{100}{20} 0.2^{20}(1-0.2)^{100-20} \approx 0.099$$

$$\mathcal{L}(20; 100, 0.15) = \binom{100}{20} 0.15^{20}(1-0.15)^{100-20} \approx 0.040$$

Binomial likelihood

The maximum likelihood occurs at $p = 0.2$



Maximum likelihood estimates

In practice, finding the MLE is not so trivial

We will use numerical optimization methods to find the MLE

Maximizing the likelihood

Let's return to our general statement for the MLE

$$\hat{\theta} = \max_{\theta} \prod_{i=1}^{n} f_{\theta}(y_i)$$

If the densities are small and/or n is large, the product will become increasingly tiny

Log-likelihood

To address this, we can make use of the logarithm function, which has 2 nice properties:

1. it's a monotonically increasing function
2. $\log(ab) = \log(a) + \log(b)$

Log-likelihood

We thereby transform our likelihood into a *log-likelihood*

$$\begin{aligned}\hat{\theta} &= \max_{\theta} \prod_{i=1}^{n} f_{\theta}(y_i) \\ &= \max_{\theta} \sum_{i=1}^{n} \log f_{\theta}(y_i)\end{aligned}$$

Maximizing the likelihood

If the data y are both independent and identically distributed, we can average over the log-likelihoods and remove the dependency on the number of observations

$$\hat{\theta} = \max_{\theta} \sum_{i=1}^{n} \log f_{\theta}(y_i)$$
$$= \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f_{\theta}(y_i)$$

Minimizing the log-likelihood

Lastly, we have been focused on minimizing functions, so we'll minimize the *negative log-likelihood*

$$\hat{\theta} = \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f_{\theta}(y_i)$$

$$\Downarrow$$

$$\hat{\theta} = \min_{\theta} -\frac{1}{n} \sum_{i=1}^{n} \log f_{\theta}(y_i)$$

Gaussian likelihood function

Let's return to the pdf for a normal distribution

$$f(y;\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

Gaussian likelihood function

Let's return to the pdf for a normal distribution

$$f(y;\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

$$\Downarrow$$

$$\begin{aligned} f(y_1,\ldots,y_n;\mu,\sigma^2) &= \prod_{i=1}^{n} f(y_i;\mu,\sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right] \end{aligned}$$

Gaussian log-likelihood function

The log-likelihood is then

$$f(y;\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

$$\Downarrow$$

$$\log f(y;\mu,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Gaussian MLE

What values of μ and σ maximize the log-likelihood?

We need to take some derivatives!

Gaussian MLE

Mean

$$\frac{\partial}{\partial\mu}\log f(y;\mu,\sigma^2) = 0 - \frac{-2n(\bar{y}-\mu)}{2\sigma^2} = 0$$

$$\Downarrow$$

$$\frac{-2n(\bar{y}-\mu)}{2\sigma^2} = 0$$

$$\Downarrow$$

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Gaussian MLE

Variance

$$\frac{\partial}{\partial \sigma} \log f(y;\mu,\sigma^2) = -\frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu) = 0$$

$$\Downarrow$$

$$\frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu)$$

$$\Downarrow$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})$$

Gaussian MLE

Variance

Recall from earlier lectures that we defined

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})$$

but our MLE is

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})$$

Gaussian MLE

Variance

Hence, our MLE for the variance is *biased* low

$$(n-1)\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \bar{y})$$

$$n\hat{\sigma}^2_{MLE} = \sum_{i=1}^{n}(y_i - \bar{y})$$

$$\Downarrow$$

$$\hat{\sigma}^2_{MLE} = \frac{n-1}{n}\hat{\sigma}^2$$

Gaussian MLE

General properties

Asymptotically, as $n \to \infty$

- estimates are *unbiased*
- estimates are normally distributed
- variance of estimate is minimized

Gaussian MLE

General properties

Invariance: if $\hat{\theta}$ is MLE of θ then $f(\hat{\theta})$ is MLE of $f(\theta)$

Gaussian MLE

Least squares estimates are MLEs

For cases where $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \Sigma)$ then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

is also the MLE for $\boldsymbol{\beta}$

Maximum likelihood estimation

Summary

Maximum likelihood estimation is much more general than least squares, which means we can use it for

- mixed effects models
- generalized linear models
- Bayesian inference