Likelihood

Michael Noonan

Biol 520C: Statistical modelling for biological data

Table of contents

1. Review
2. Likelihood
3. Maximum likelihood
4. Regression as a problem of MLE

Review

Probability theory & Bayes' Theorem

Last lecture we covered some of the basic rules of probability theory including joint and conditional probabilities and Bayes' theorem.

We learned that the conditional probability of A given B occurred can be formally written as:

$$\mathrm{Pr}\{\mathrm{A}|\mathrm{B}\} = \mathrm{Pr}\{\mathrm{A,B}\}/\mathrm{Pr}\{\mathrm{B}\}$$

and that the joint probability of A and B can be formally written as:

$$\mathrm{Pr}\{\mathrm{A,B}\} = \mathrm{Pr}\{\mathrm{A}\cap\mathrm{B}\} = \mathrm{Pr}\{\mathrm{A}\}\mathrm{Pr}\{\mathrm{B}\}$$

We also learned how the basic rules of probability get us to the mathematical description of Bayes' Theorem.

And I told you that this was important for fitting a model to data.

Likelihood

Probability distributions

Last lecture we saw that $\Pr\{A|B\}$ represents the conditional probability of observing event ‘A’ given event ‘B’, but A and B are not restricted to being events.

For example, we can extend this to indicate that the probability of observing data Y_i given parameter value p is $\Pr\{Y_i|p\}$.

The subscript i indicates that there are multiple possible outcomes, but only one parameter value p.

Probability distributions cont.

Let's say I've gone out and counted things.

Count data often follow a Poisson distribution, so $\Pr\{Y_i|p\}$ can be written as:

$$\Pr\{Y_i = k|\,\mathrm{rate\,parameter}\ = \lambda\} = \frac{\lambda^k e^{-\lambda}}{k!}$$

In other words, we have a mathematical description for saying "What is the probability of observing our data (Y_i) given our hypothesis (rate parameter $= \lambda$)?"

Likelihoods vs. probabilities

This framework provides us with the tools we need to quantify the probability of our observation given some hypothesis.

In most biological situations, however, all we have are the data and we don't know what the underlying model is. What we're really interested is asking "Given these data, how likely are the different hypotheses".

This is where the concept of likelihoods comes in, and we write this:

$$\mathcal{L}(\text{hypothesis }|\text{data }) \text{ or } \mathcal{L}(p_m|Y)$$

Note how here the data are not subscripted (we only observed one outcome), but there are multiple possible parameter values p_m.

Likelihoods vs. probabilities cont.

The key distinction between likelihoods and probabilities is that with probabilities the hypothesis is *known*, but the data are *unknown* whereas with likelihoods the data are *known* but the hypothesis is *unknown*.

If we assume that the likelihood of the data given the hypothesis is *proportional* to the probability, we can write:

$$\mathcal{L}(p_m|Y) \quad = \quad c \quad \Pr\{Y|p_m\}$$

Because we're typically interested in relative likelihoods (not exact values), the proportionality constant, c, can be set to 1

Likelihoods vs. probabilities cont.

Ok, cool, but so what?

Likelihoods vs. probabilities cont.

If $\mathcal{L}(p_m|Y) \propto \Pr\{Y|p_m\}$, and we have some data, and if we make some distributional assumptions, we have a way of calculating the likelihood of specific parameter values, and we can generalise this to any number of parameters and any distribution!

For example, if we come back to those 2 crows I counted earlier we can now ask (and answer!) what's the more likely λ, 4, or 6?

$$\mathcal{L}(\lambda = 4|2) = \frac{4^2 e^{-4}}{2!} \approx 0.15 \qquad \mathcal{L}(\lambda = 6|2) = \frac{6^2 e^{-6}}{2!} \approx 0.045$$

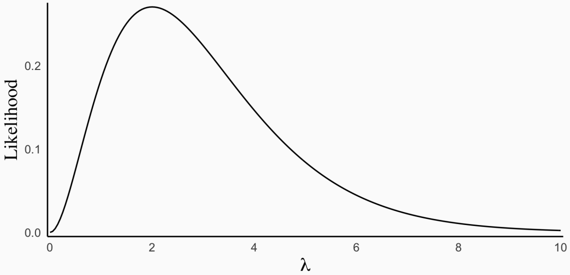
Remember that the Poisson distribution has a PMF given by $\frac{\lambda^k e^{-\lambda}}{k!}$

Maximum likelihood

Maximum Likelihood

Likelihood gives us a framework of estimating the *MOST* likely value of λ by systematically checking every possible value of $\Pr\{2|\lambda_i\}$.

For our 2 crows, a plot of the likelihood as a function of λ looks like this:



The value of λ that maximises the likelihood is the maximum likelihood estimate (MLE) of our parameter value

Multiple observations

In most cases we probably have more than 1 data point though...

So let's say I go out another day and count 4 crows. Now my data are $\{2,4\}$, and our question is $\mathcal{L}(\lambda_1|2,4)$.

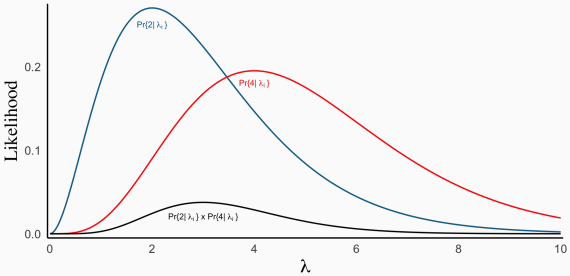
To answer this, we need to remember our probability theory basics.

If $\mathcal{L}(p_m|A) \propto \Pr\{A|p_m\}$, then $\mathcal{L}(p_m|A,B) \propto \Pr\{A,B|p_m\}$, so we now know we need to calculate $\Pr\{A,B|p_m\}$

If we assume our observations are independent, then $\Pr\{A,B|p_m\} = \Pr\{A|p_m\} \times \Pr\{B|p_m\}$

Multiple observations cont.

For our counts of 2 and 4 crows, a plot of the likelihood as a function of λ looks like this:



Here, $\hat{\lambda} = 3$ is the maximum likelihood estimate (MLE)

Multiple observations cont.

More generally, if we were to keep collecting data, we would keep multiplying out probabilities to estimate λ

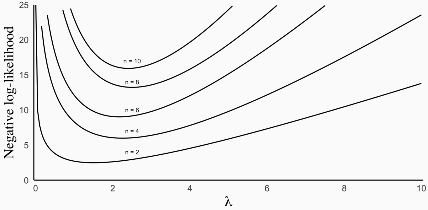
$$\mathcal{L}(\lambda|x_i) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Log-likelihoods

Likelihoods tend to be very small numbers, so by convention we work with log-likelihood in practice as it makes the math/computations easier.

And we actually minimise the negative log-likelihood (but functionally the results are the same as maximising likelihoods)

In most cases, as the sample size increases, the negative log-likelihood function becomes increasingly peaked around its maximum

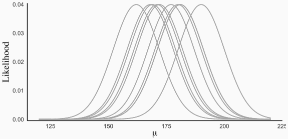


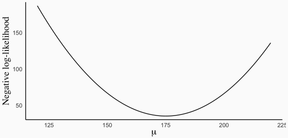
MLE in action

Let's say I go out and measured the height of ten people. I come back with {171,168,180,190,169,172,162,181,177,181} (in cm). If we assume heights are normally distributed with a σ of 10 cm, we can estimate the mean height, μ, using MLE and the general form of the Gaussian distribution:

$$\mathcal{L}(\mu|x_i) = \prod_{i=1}^{10} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2}$$

Plugging in data the resulting likelihood functions look like this:

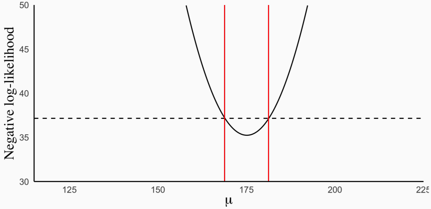




MLE in action cont.

The MLE estimate of the mean is 175.1, which is exactly the same as the sample mean, so why go through all the extra effort?

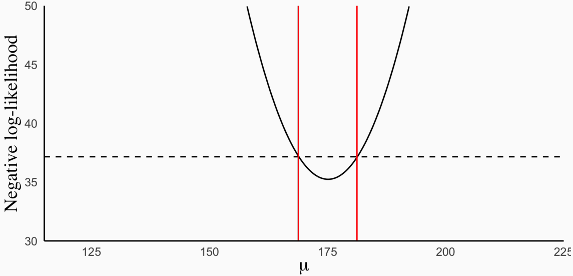
In the likelihood framework, we can use the likelihood profile to identify the 95% confidence intervals on our estimated parameter. E.g., a simple rule is to place the bounds within 1.92 of the minimum log-likelihood.



So we can say that $\mu = 175.1$ with 95% CIs of $\sim$ 169 — 181

MLE and CIs.

Earlier I mentioned that as the sample size increases, the negative log-likelihood becomes increasingly peaked around its maximum. What would this do to the CIs?



More data narrows the CIs (and vice versa, less data increases the amount of uncertainty in the MLE)

Regression as a problem of MLE

MLE and fitting a line to data

We started our detour into probability theory and maximum likelihood because we learned that the least-squares approach didn't provide any way of understanding how stochasticity entered into a system.

We've now picked up enough of the basics to go back to our linear regression problem and fit a line to some data as probabilists.

Regression as a problem of MLE

Our model is of the form:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

y_i & x_i are our data,

β_0 & β_1 are our unknown parameters,

and ε_i is our Gaussian distributed error with a mean of 0 and variance of σ^2, whose PDF is given by:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Regression as a problem of MLE cont.

Given our dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, and assuming independence in y_i, the likelihood function be written:

$$\mathcal{L}(x_i; \beta_0, \beta_1, \sigma^2 | y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y_i - \beta_0 + \beta_1 x_i}{\sigma})^2}$$

For convenience, we work with negative log-likelihoods, which changes this to:

$$\mathcal{L}(x_i; \beta_0, \beta_1, \sigma^2 | y_i) = n(\log(\sigma) + \tfrac{1}{2}\log(2\pi)) + \tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Regression as a problem of MLE cont.

We could manually plug in all of the possible values of β_0, β_1, and σ^2, but that would be a lot of work.

After some math that we won't go over, we get the following 3 estimators:

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Stochasticity & CIs

The MLE estimators for the slope and the intercept exactly match the least squares estimators, and σ^2 is the mean squared error.

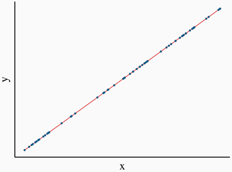
So what did we gain by making a Gaussian-noise assumption and estimating the parameters via maximum likelihood?

i) A description of the system's stochastic component (which allows us to make more realistic predictions)

ii) A framework for placing confidence intervals on our parameter estimates using the likelihood profiles (which allows us to make formal statistical inference on the parameters)

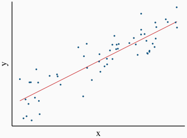
Stochasticity and predictions

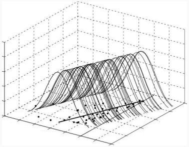
With a purely deterministic model, we can understand the mechanisms underlying a system, but our predictions inherently lack stochasticity and the outcome is always the same.



But real systems are full of stochasticity, so the predictions of purely deterministic models are almost certainly going to be wrong.

With a stochastic component, outcomes are variable and models provide a distribution of the values that y_i can be expected to take





Source: Zuur et al. 2009

Stochasticity and predictions example

For example, let's say we know that the wingspan (in mm) a species of insect is proportional to its mass (in g), with a slope of 5 and intercept of 10. How wide does this deterministic model predict our species' wings will be when $x = 2g$?

$y = 10 + 5 \times 2 = 20mm$

In other words, this model predicts our species of insect will have a wingspan of 20 mm when their mass is 2g, not 20.1, not 20.00001, but exactly 20.

How informative is this? What if we measure an insect with weight = 2 and wingspan = 21? What does this tell us about our model?

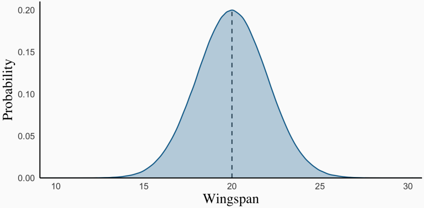
Stochasticity and predic. examp. cont.

What if we also know that $\varepsilon_i \sim \mathcal{N}(0, 2)$?

$y_i = 10 + 5x_i + \varepsilon_i$

$y = 10 + 5 \times 2 + \varepsilon_i = 20 + \varepsilon_i$

$y = \mathcal{N}(20, 2)$



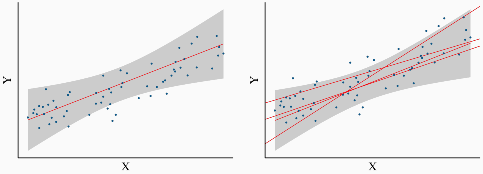
Now we have a prediction of the distribution of all the values that y_i can be expected to take. How informative is this?

A note on predictions

Confidence intervals tell us how well we have estimate a parameter of interest, such as a mean or regression coefficient.

$$\hat{y} \pm t_{crit} \times \sigma\sqrt{\frac{1}{n}}$$

For linear regression, 95% CIs tell us that there is a 95% probability that the true linear regression line of the population will lie within the confidence interval of the regression line calculated from the data.



What happens to our CIs when $n \rightarrow \infty$?

A note on predictions cont.

In the limit where $n \to \infty$, we will have perfect information on the population so our CIs collapse to the population value, but our predictions still need to account for stochasticity in the individual values.

To account for both the uncertainty in estimating the parameters, **plus** the random variation of the individual values we use prediction intervals.

$$\hat{y} \pm t_{crit} \times \sigma \sqrt{1 + \frac{1}{n}}.$$

What happens to prediction intervals when $n \to \infty$? Which will be wider, confidence intervals, or prediction intervals?

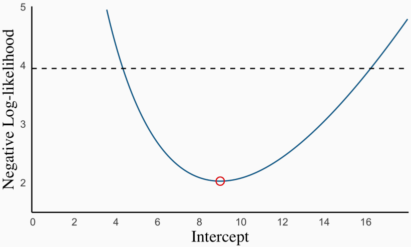
MLE and Confidence Intervals

Without a measure of (un)certainty, it is impossible to tell if two numbers are **meaningfully** different from one another.

The same is true for model parameters. We are estimating their values based in some amount of data, which represents a subset of the infinite number of possibilities we could have actually observed.

For example, if model fittings suggest an intercept of 9, is this meaningfully different from 0?

The answer depends on the amount of information in our data, the (un)certainty in our estimate, and the shape of our likelihood profile.


Negative Log-likelihood
Intercept