

02-712 Week 12

Biological Modeling and Simulation

Aidan Jan

November 11, 2025

Regression

Suppose we want to fit a straight line to our data (represented by points on the cartesian plane). We can do that by posing this as a gradient descent problem.

- Write the equation for our line (it's probably $y = mx + b$)
- We can calculate error by subtracting the inferred y -value from the line from the true y of the data point.
- Sum the squares of the error. We do squares to prevent negative errors from cancelling positive errors.

$$SE = (y_1 - (mx_1 + b))^2 + (y_m - (mx_m + b))^2$$

- From here, our goal is to find (m, b) such that it minimizes error. To do this, we take the derivative of the squared-error equation in terms of m and b .
- By expanding, then taking the derivative, we get the system

$$\begin{aligned}\frac{\partial SE}{\partial m} &= 0 \implies m\bar{x}^2 + b\bar{x} = \bar{xy} \\ \frac{\partial SE}{\partial b} &= 0 \implies m\bar{x} + b = \bar{y}\end{aligned}$$

- What can we interpret from these equations? (the ‘bar’ symbol represents average.)
- The second equation tells us that the average point of all the points must be on the best fit line.
- The first equation is the same, but less intuitive, which is the point $\left(\frac{\bar{x}^2}{\bar{x}}, \frac{\bar{xy}}{\bar{x}}\right)$ must be on the line.

Example: Maximum Likelihood Estimation (MLE)

Maximum Likelihood is a probabilistic method that determines values for the parameters of the model. The parameter values are found such that they maximize the likelihood that the process described by the model produces the observed data.

Suppose we have observations that are sampled (for this example, let's assume from a Gaussian). Therefore,

$$P(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Let our data points be 9, 9.5, 11. In this case, we want to maximize

$$P(\text{data points}, \mu, \sigma) = P_1 \cdot P_2 \cdot P_3$$

The goal is to find (μ, σ) that maximizes the probability above.

Example: Allele Densities

Suppose we have a population of animals with two different alleles of some trait, A and a . Let the population have N individuals. The probability that a individual is a mutant (allele = A) is p . How do we find p ?

- Well... we can always count the number of A 's in our sample and just use that as a guess for the population. But can we do better?

We can pose this as a maximum likelihood estimation problem. First, define a dummy variable.

$$x_i = \begin{cases} 1 & A \\ 0 & a \end{cases}$$

Now, let's define a function, f , that tells us the probability of getting our observations (under our observations.)

$$f(x_i|p) = p^{x_i} \cdot (1-p)^{1-x_i}$$

We use the binomial form here, since each variable follows a Bernoulli distribution (either mutant or not). This equation works for one individual, so now we can expand it to all individuals.

$$f(x_1, \dots, x_n|p) = p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n}$$

This can be written as a product:

$$= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$$

Now, we want to find $\frac{\partial L}{\partial p} = 0$ since we want to maximize the likelihood. It turns out that this function sucks to take the derivative of, so instead, we take the log of the function.

$$\ell = \log \mathcal{L} = \log \left(\prod_{i=1}^N p^{x_i}(1-p)^{(1-x_i)} \right)$$

This now converts into a sum of logs, and with some algebra to simplify, we will eventually get to

$$= N\bar{x}\log(p) + N(1-\bar{x})\log(1-p)$$

We can now take the derivative:

$$\frac{\partial \log(\mathcal{L})}{\partial p} = \frac{\partial \ell}{\partial p} = \frac{N\bar{x}}{\hat{p}} - \frac{N(1-\bar{x})}{1-\hat{p}} = 0$$

Now, we can simplify:

$$\begin{aligned} 0 &= \frac{N\bar{x}}{\hat{p}} - \frac{N(1-\bar{x})}{1-\hat{p}} \\ \frac{N(1-\bar{x})}{1-\hat{p}} &= \frac{N\bar{x}}{\hat{p}} \\ \hat{p}(1-\bar{x}) &= \bar{x}(1-\hat{p}) \\ \hat{p} &= \bar{x} \end{aligned}$$

It turns out that our intuition at the beginning was correct.

Least Squares as Maximum Likelihood

$$y_i = \beta x_i + \varepsilon_i$$

We let (x_i, y_i) be our data point and ε_i be our error term, where $\varepsilon_i \sim N(0, \sigma^2)$. We can solve this equation in terms of our error, which gives:

$$\varepsilon_i = y_i - \beta x_i \sim N(0, \sigma^2)$$

Therefore,

$$p(x_i|\beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\beta x_i)^2}{2\sigma^2}}$$

The likelihood for points in our sample is

$$\mathcal{L} = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\beta x_i)^2}{2\sigma^2}}$$

To solve this, we use the same method as before, taking the log of this, then the derivative.