

## Maximum Likelihood

---

Consider IID random samples  $X_1, X_2, \dots, X_n$  where  $X_i$  is a sample from the density function  $f(X_i|\theta)$ . We are going to introduce a new way of choosing parameters called Maximum Likelihood Estimation (MLE). We want to select that parameters ( $\theta$ ) that make the observed data the most likely. *Note that we are now using notation that shows that the density of  $X$  depends on its parameters,  $\theta$ .*

First we define the likelihood of our data give parameters  $\theta$ :

$$L(\theta) = \prod_{i=1}^n f(X_i|\theta)$$

This is the probability of all of our data. It evaluates to a product because all  $X_i$  are independent. Now we chose the value of  $\theta$  that maximizes the likelihood function. Formally  $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta)$ .

A cool property of  $\operatorname{argmax}$  is that since  $\log$  is a monotone function, the  $\operatorname{argmax}$  of a function is the same as the  $\operatorname{argmax}$  of the  $\log$  of the function! That's nice because logs make the math simpler. Instead of using likelihood, you should instead use log likelihood:  $LL(\theta)$ .

$$LL(\theta) = \log \prod_{i=1}^n f(X_i|\theta) = \sum_{i=1}^n \log f(X_i|\theta)$$

To use a maximum likelihood estimator, first write the log likelihood of the data given your parameters. Then chose the value of parameters that maximize the log likelihood function.  $\operatorname{Argmax}$  can be computed in many ways. Most require computing the first derivative of the function.

### Bernoulli MLE Estimation

Consider IID random variables  $X_1, X_2, \dots, X_n$  where  $X_i \sim \operatorname{Ber}(p)$ . First we are going to write the PMF of a Bernoulli in a crazy way: The probability mass function  $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$ . Wow! Whats up with that? First convince yourself that when  $X_i = 0$  and  $X_i = 1$  this returns the right probabilities. We write the PMF this way because its derivable.

Now let's do some MLE estimation:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} \\ LL(\theta) &= \sum_{i=1}^n \log p^{X_i}(1-p)^{1-X_i} \\ &= \sum_{i=1}^n X_i(\log p) + (1-X_i)\log(1-p) \\ &= Y \log p + (n-Y)\log(1-p) \end{aligned} \quad \text{where } Y = \sum_{i=1}^n X_i$$

Great Scott! Now we simply need to chose the value of  $p$  that maximizes our log-likelihood. One way to do that is to find the first derivative and set it equal to 0.

$$\begin{aligned} \frac{\delta LL(p)}{\delta p} &= Y \frac{1}{p} + (n-Y) \frac{-1}{1-p} = 0 \\ \hat{p} &= \frac{Y}{n} = \frac{\sum_{i=1}^n X_i}{n} \end{aligned}$$

All that work and we get the same thing as method of moments and sample mean...

## Normal MLE Estimation

Consider IID random variables  $X_1, X_2, \dots, X_n$  where  $X_i \sim N(\mu, \sigma^2)$ .

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(X_i | \mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \\ LL(\theta) &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \\ &= \sum_{i=1}^n \left[ -\log(\sqrt{2\pi\sigma}) - \frac{1}{2\sigma^2} (X_i - \mu)^2 \right] \end{aligned}$$

If we chose the values of  $\hat{\mu}$  and  $\hat{\sigma}^2$  that maximize likelihood, we get:  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$  and  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2$ .

## Linear Transform Plus Noise

Assume that  $Y = \theta X + Z$  where  $Z \sim N(0, \sigma^2)$  and  $X$  is an unknown distribution. The equations imply that  $Y|X \sim N(\theta X, \sigma^2)$ . Choose a value of  $\theta$  that maximizes the probability of the data:  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ .

We approach this problem by finding a function for the log likelihood of the data given  $\theta$ . Then we find the value of  $\theta$  that maximizes the log likelihood function. To start, use the PDF of a Normal to express the probability of  $Y|X, \theta$ :

$$f(Y_i | X_i, \theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}}$$

Now we are ready to write the likelihood function, then take its log to get the log likelihood function:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f(Y_i, X_i, \theta) && \text{Let's break up this joint} \\ &= \prod_{i=1}^n f(Y_i | X_i, \theta) f(X_i) && f(X_i) \text{ is independent of } \theta \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} f(X_i) && \text{Substitute in the definition of } f(Y_i | X_i) \end{aligned}$$

$$\begin{aligned} LL(\theta) &= \log L(\theta) \\ &= \log \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} f(X_i) && \text{Substitute in } L(\theta) \\ &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} + \sum_{i=1}^n \log f(X_i) && \text{Log of a product is the sum of logs} \\ &= m \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^m (Y_i - \theta X_i)^2 + \sum_{i=1}^n \log f(X_i) \end{aligned}$$

Remove constant multipliers and terms that don't include  $\theta$ . We are left with trying to find a value of  $\theta$  that maximizes:

$$\begin{aligned} \hat{\theta} &= \operatorname{argmax}_{\theta} - \sum_{i=1}^m (Y_i - \theta X_i)^2 \\ &= \operatorname{argmin}_{\theta} \sum_{i=1}^m (Y_i - \theta X_i)^2 \end{aligned}$$

This result says that the value of  $\theta$  that makes the data most likely is one that minimizes the squared error of predictions of  $Y$ . We will see in a few days that this is the basis for linear regression.