University of BRISTOL

WEEK 3

Maximum likelihood estimation

Felipe Campelo

bristol.ac.uk

Partly adapted from Montgomery and Runger's *Applied Statistics and Probability for Engineers*

In this lecture…

In this lecture, we’ll cover how the method of maximum likelihood can be used to construct estimators.

The estimation problem

When we use a probability distribution to model a certain data-generating phenomenon based on some data, we are performing **statistical modelling**.

Given a probability distribution $\mathrm{P}(X|\theta)$ (i.e., parameterised by θ) and a sample $X_1, X_2, \dots, X_N$, we would like a general strategy for finding (near) optimal estimators $\widehat{\Theta}$ for population parameters θ – i.e., to *find the estimate* $\hat{\theta}$ *so that the model* $P(X|\theta)$ *best fits the data.*

We have seen some intuitive examples (sample mean, sample variance, sample proportion etc.) but is there a more general method?

“*Essentially,
all models are wrong,
but some are useful.*”
George E.P. Box, 1987

The likelihood function

Given a probability distribution $P(X|\theta)$ and a sample $\boldsymbol{X} = \{X_1, \ldots, X_N\}$, we want to have an estimators $\hat{\Theta}$ that provides good estimates of θ.

The ***likelihood function*** $L(\theta|\boldsymbol{X})$ (sometimes written as $L(\theta;\boldsymbol{X})$) is a function that maps each value of θ to a single number between 0 and 1 that quantifies the ***goodness-of-fit*** between $\mathrm{P}(X|\theta)$ and the sample $\boldsymbol{X}$. Given a sample $\boldsymbol{X}$, the function is defined as:

Discrete random variables: $L(\theta|\mathbf{X}) \coloneqq \prod_{i=1}^{N} p_\theta(X_i)$,
with p_θ being the probability mass function.

Continuous random variables: $L(\theta|\mathbf{X}) \coloneqq \prod_{i=1}^{N} f_\theta(X_i)$,
with p_θ being the probability density function.

Example: a discrete random variable

Suppose you have an i.i.d. sample $\boldsymbol{X} = \{X_1, \dots, X_N\}$ that you believe come from a data-generating process following a Bernoulli distribution $Bern(X|q_0)$, with unknown parameter q_0.

Since the sample is i.i.d., every observation has a pmf given by

$$p(x|q) = q^x(1-q)^{1-x}, \text{for x} \in \{0,1\}$$

The likelihood function is therefore given by:

$$L(q|\mathbf{X}) \coloneqq \prod_{i=1}^{N} p(X_i) = \prod_{i=1}^{N} q^{X_i}(1-q)^{1-X_i} = q^{\Sigma X_i}(1-q)^{(n-\Sigma X_i)}$$

Example: a continuous random variable

Suppose you have an i.i.d. sample $\boldsymbol{X} = \{X_1, \dots, X_N\}$ from from a process known to follow a Gaussian distribution $\mathcal{N}(X|\mu, \sigma^2)$, with unknown mean and variance.

Since the sample is i.i.d., every observation has a pdf given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \text{for x} \in \mathbb{R}$$

The likelihood function is therefore given by:

$$L(\mu, \sigma^2|\mathbf{X}) := \prod_{i=1}^{N} f(X_i|\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum\left(\frac{X_i-\mu}{\sigma}\right)^2\right)$$

Maximum likelihood estimation

Since the likelihood function provides a normalised way to quantify how well a given probability distribution fits a given sample given a value of θ, an intuitive way to find the best estimator is to solve the *optimisation problem*:

$$Find\ \hat{\theta} = \underset{\theta \in \Theta}{\arg max}\, L(\theta|\boldsymbol{X})$$

i.e., ***the value of $\widehat{\theta}$ that maximises the likelihood of the model, given the data.***

In some cases, this problem can be solved analytically (taking the derivative of the function in relation to the parameter and equating it to zero). Note that it is usually easier to maximise the log-likelihood function instead of $L(\theta|\mathbf{X})$.

$$l(\theta|\mathbf{X}) = logL(\theta|\mathbf{X}) = \begin{cases} \sum_{i=1}^{N} p_\theta(X_i), & \text{if RV is discrete} \\ \sum_{i=1}^{N} f_\theta(X_i), & \text{if RV is continuous} \end{cases}$$

Example: a discrete random variable

Remember the Bernoulli likelihood function,

$$L(q|\mathbf{X}) \coloneqq\ q^{\sum X_i}(1-q)^{(n-\sum X_i)}$$

The corresponding log-likelihood is

$$\mathrm{l}(q|\mathbf{X}) \coloneqq \log(q)\sum X_i + (n-\sum X_i)\log(1-\mathrm{q})$$

Taking the derivative w.r.t. q and equating to zero results in

$$\frac{dl}{dq} = \frac{\sum X_i}{q} - \frac{n-\sum X_i}{1-q} = 0 \rightarrow \hat{q}_{MLE} = \frac{\sum X_i}{n} = \bar{X}$$

i.e., the MLE estimator of the probability parameter q is the sample mean.

Example: a continuous random variable

Remember the Gaussian likelihood function,

$$L(\mu, \sigma^2 | \mathbf{X}) := \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum\left(\frac{X_i - \mu}{\sigma}\right)^2\right)$$

The corresponding log-likelihood is

$$\mathrm{l}(\mu, \sigma^2 | \mathbf{X}) := -n \log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2}\sum\left(\frac{X_i - \mu}{\sigma}\right)^2$$

Since there are two parameters to estimate, we need to take partial derivatives w.r.t. each and equate both to zero.

Equating the partial derivatives with zero,

$$\partial l \,/ \partial\mu = \sum \frac{X_i - \mu}{\sigma} = 0$$

$$\frac{\partial l}{\partial\sigma} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(X_i - \mu)^2 = 0$$

and solving the system above results in:

$$\hat{\mu}_{MLE} = \frac{\sum X_i}{n}$$

$$\widehat{\sigma^2}_{MLE} = \frac{\sum(X_i - \bar{X})^2}{n}$$

Example: a continuous random variable

For the Gaussian distribution,

$$\hat{\mu}_{MLE} = \frac{\sum X_i}{n} = \bar{X} \qquad \widehat{\sigma^2}_{MLE} = \frac{\sum (X_i - \bar{X})^2}{n}$$

Notice that the MLE estimator of the mean is the MVUE, *but the one for variance isn't* because it is not an unbiased estimator. Remember, the MVUE for the variance has a slightly different denominator[1]:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

The MLE estimator is, however, still close enough in most practical situations.

[1] https://web.ma.utexas.edu/users/mks/M358KInstr/SampleSDPf.pdf

Further reading

DC Montgomery, GC Runger, *Applied statistics and probability for engineers, 5th ed.* **[Chapter 7.4.2]**

- *Read the chapter and try to solve the worked examples by yourself*.