Fundamentals of Data Science

Maximum likelihood

Ramesh Johari

The likelihood function

Estimating the parameter

This lecture develops the methodology behind the *maximum likelihood* approach to parameter estimation.

Basic idea:

- In guessing what the parameters are, we first ask: what is the chance of seeing the data we observe, given a particular value of the parameter?
- We then pick the parameter values that maximize this chance.

Example 1: Flipping a biased coin

Suppose we flip a coin n times, and get observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

If the bias on the coin was q, what is the probability we get data $\mathbf{Y}$?

This is the *likelihood* of $\mathbf{Y}$ given q:

$$f(\mathbf{Y}|q) = \prod_{i=1}^{n} q^{Y_i}(1-q)^{1-Y_i}.$$

Example 2: Linear normal population model

Suppose that you are given p-dimensional covariate vectors $\mathbf{X}_i$, $i = 1, \ldots n$, together with corresponding observations Y_i, $i = 1, \ldots, n$.

Suppose you assume the *linear normal population model* with *i.i.d. errors*:

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and the ε_i are i.i.d.

What are the parameters? What is the likelihood?

Example 2: Linear normal population model

Parameters are $\boldsymbol{\beta}$ and σ^2.

In deriving likelihood, we will typically treat $\mathbf{X}$ as *given*; i.e., we focus on inference about the relationship between $\mathbf{X}$ and $\mathbf{Y}$, rather than the distribution of $\mathbf{X}$.

Likelihood is probability density of seeing $\mathbf{Y}$, given parameters and $\mathbf{X}$:

$$f(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left(-\frac{(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2}{2\sigma^2} \right).$$

In general: Parametric likelihood

In general, suppose there is some process that generates data $\mathbf{Y}$.[1]

Suppose each choice of a parameter vector $\boldsymbol{\theta}$ gives rise to a conditional pmf (or pdf) $f(\mathbf{Y}|\boldsymbol{\theta})$.

This is the probability (or density) of seeing $\mathbf{Y}$, given parameters $\boldsymbol{\theta}$.

We call this the *likelihood* of $\mathbf{Y}$ given $\boldsymbol{\theta}$.

[1] As noted on the previous slide, in the case of regression, we treat $\mathbf{X}$ as given and look at the process that generates $\mathbf{Y}$ from $\mathbf{X}$.

Log likelihood

It is often easier to work with logs when taking likelihoods (not least because multiplying many small probabilities together can cause numerical instability).

We call this the *log likelihood function* (LLF).

Example 1 (biased coin):

$$\log f(\mathbf{Y}|q) = \sum_{i=1}^{n} Y_i \log q + (1 - Y_i) \log(1 - q).$$

Example 2 (linear normal population model):

$$\log f(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2.$$

Maximum likelihood

Maximizing the log likelihood

The *maximum likelihood estimate* (MLE) is the parameter value that maximizes the likelihood.

It is found by solving the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & f(\mathbf{Y}|\boldsymbol{\theta}) \quad (\text{or } \log f(\mathbf{Y}|\boldsymbol{\theta})) \\ \text{over} \quad & \text{feasible choices of } \boldsymbol{\theta}. \end{aligned}$$

We denote the resulting solution by $\hat{\boldsymbol{\theta}}_{\text{MLE}}$.

Example 1: Flipping a biased coin

Suppose we maximize the log likelihood. The resulting solution is:

$$\hat{q}_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}.$$

This is fairly reasonable: we estimate q by the mean number of successes.

Example 2: Linear normal model, known σ^2

Suppose that σ^2 is known, so the goal is to estimate $\boldsymbol{\beta}$. Returning to LLF, our problem is equivalent to choosing $\hat{\boldsymbol{\beta}}$ to minimize:

$$\text{minimize} \quad \sum_{i=1}^{n} (Y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})^2.$$

In other words: the OLS solution *is* the MLE estimate of the coefficients!

Example 2: Linear normal model, known σ^2

Let's pause to reflect on what we've found here.

When we first discussed OLS, we did so from a purely algebraic view — no assumptions.

Example 2: Linear normal model, known σ^2

Let's pause to reflect on what we've found here.

When we first discussed OLS, we did so from a purely algebraic view — no assumptions.

When we discussed prediction, we showed that as long as the population model is linear and the ε_i are *uncorrelated* with $\mathbf{X}_i$, and we include the right set of covariates in our regression, then OLS is unbiased, and has minimum variance among unbiased linear estimators (Gauss-Markov theorem).

Example 2: Linear normal model, known σ^2

Let's pause to reflect on what we've found here.

When we first discussed OLS, we did so from a purely algebraic view — no assumptions.

When we discussed prediction, we showed that as long as the population model is linear and the ε_i are *uncorrelated* with $\mathbf{X}_i$, and we include the right set of covariates in our regression, then OLS is unbiased, and has minimum variance among unbiased linear estimators (Gauss-Markov theorem).

Now, we have shown that in addition if we assume the ε_i are i.i.d. normal random variables, then OLS is the maximum likelihood estimate.

Example 2: Linear normal model, unknown σ^2

What happens if σ^2 is *unknown*?

The MLE for $\boldsymbol{\beta}$ remains unchanged (the OLS solution), and the MLE estimate for σ^2 is:

$$\hat{\sigma}^2_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2.$$

This is intuitive: the sum of squared residuals is an estimate of the variance of the error.

Example 3: Logistic regression

The logistic regression parameters are found through *maximum likelihood*.

The (conditional) likelihood for $\mathbf{Y}$ given $\mathbf{X}$ and parameters $\boldsymbol{\beta}$ is:

$$\mathbb{P}(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^{n} g^{-1}(\mathbf{X}_i\boldsymbol{\beta})^{Y_i}(1 - g^{-1}(\mathbf{X}_i\boldsymbol{\beta}))^{1-Y_i}$$

where $g^{-1}(z) = \exp(z)/(1 + \exp(z))$.
Let $\hat{\boldsymbol{\beta}}_{\mathsf{MLE}}$ be the resulting solution; these are the *logistic regression coefficients*.

Example 3: Logistic regression [$*$]

Unfortunately, in contrast to our previous examples, maximum likelihood estimation does not have a closed form solution in the case of logistic regression.

However, it turns out that there are reasonably efficient iterative methods for algorithmically computing the MLE solution.

One example is an algorithm inspired by weighted least squares, called *iteratively reweighted least squares*. This algorithm iteratively updates the weights in WLS to converge to the logistic regression MLE solution. See [AoS], Section 13.7 for details.