The Logic of Maximum Likelihood Estimation

Jan Rovny

February 2, 2020

What is Maximum Likelihood Estimation (MLE)

- MLE is a unified method of statistical estimation.
  - Through one logic it provides a framework for estimation and hypothesis testing.
- MLE is theoretically grounded in that it requires an explicit model of the data generation process.
  - We must explicitly assume the distribution which gave rise to our dependent variable.
- MLE is extremely versatile.
  - It allows estimation of simple linear models, as well as complex models which are non-linear in the parameters.
  - We can estimate models with binary, ordinal or nominal dependent variables, as well as many other classes of models, using MLE.
- MLE has desirable asymptotic properties.
- Consequently, many different classes of MLE models are widely used in social sciences.

The History of MLE

- MLE statistical theory was developed by a British geneticist and statistician Robert A. Fisher between 1912 and 1922.
- Its application, however, had to wait until the 1980s and 1990s, for one simple reason: most ML estimates cannot be found analytically. They are too complicated to calculate.
  - MLE requires taking the first derivative of the log-likelihood function, which is easy in linear models, but becomes analytically intractable with complex functions.
  - Solutions have to be found through numerical optimization methods, which essentially require sufficient computing power.
- MLE thus become widely used only once our computers become powerful enough to solve the models.

Probability Theory

- MLE is rooted in probability theory, but in reverse.
- **In probability theory:**
  - We know the parameter value, and we try to predict particular data.
  - For example, we know that the probability of getting *heads* on a flip of a fair coin is $\pi = 0.5$. We can then ask ourselves how many *heads* we are **likely** to observe in 10 flips.
  - The answer is of course 5, but if you try it right now, you might get another number.
  - 5 *heads* in 10 flips of a fair coin is simply the **most likely** outcome.

Maximum Likelihood Logic

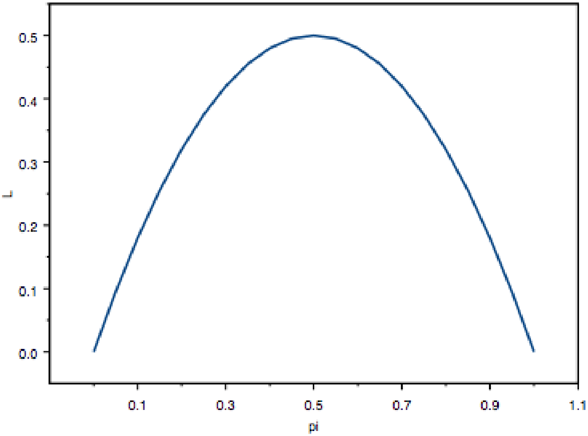
- **In statistical estimation**, we are in the reverse situation:
  - We know the data, but we do not know the parameter value (the estimate) that produced it
- The MLE question thus stands: **given my data, what is the value of the parameter that most likely produced the data?**
  - What value of π is **most likely** behind my observing 5 *heads* in 10 flips of a coin?
  - The answer is of course $\pi = 0.5$, but again, this is only the **most likely** value.
  - Given random error, it is possible to observe 5 *heads* in 10 flips even with an unfair coin where $\pi = 0.4, 0.6$, or even 0.3.

The Likelihood Function

- You can see from the previous example that different values of π *could have* produced our 5 *heads* in 10 flips.
- Only one value of π, however, is **most likely**.
- We can thus think of a function that would describe the probability (or likelihood) that different parameter values produced our data.
- This is the **likelihood function**.
- The value which *maximizes* the likelihood function is the **Maximum Likelihood Estimator**.

The Likelihood Function

The likelihood function for π, given 5 *heads* in 10 flips of a coin:



The Likelihood Function

More formally:

Let $f(\mathrm{y}|\theta)$ be the Probability Density Function (PDF) consisting of a single parameter θ, and let D denote the observed data consisting of n independent observations.

$$\begin{aligned}
\Pr(D) &= f(y_1, y_2, ..., y_n|\theta) && (1)\\
&= f(y_1|\theta)f(y_2|\theta)\dots f(y_n|\theta) && (2)\\
&= \prod f(y_i|\theta) && (3)\\
&= \mathcal{L}(\theta|D) && (4)
\end{aligned}$$

- (2) follows from the independence of observations. (The probability of two independent observations is the product of their individual probabilities).
- The step from (3) to (4) is the 'reversal' from probability (where we explain outcomes with parameters) to likelihood (where we explain parameters with outcomes or data).

Distributional Assumptions

- Notice that the likelihood function is based on a Probability Density Function which gave rise to our data.
- Thus, whenever we do MLE, we must make a **distributional assumption** about our data.
  - Based on our knowledge and theoretical expectations, we – as researchers – need to decided what distribution is behind our data (binomial, poisson, normal...).
  - Consequently, it is necessary to think through the theoretical background to a phenomenon, and the distributional implications that it has.
  - This makes MLE a more theoretically rich method than OLS or GLS, which are essentially "data fitting."
  - *"All knowledge is a result of theory – we buy information with assumptions"* (Coombs 1976:5)

Log-Likelihood

To make life easier for ourselves and our computers, we take the natural logarithm of our likelihood function before maximization.

$$\begin{aligned} ln\mathcal{L}(\theta|D) &= ln[\prod f(y_i|\theta)] &(5)\\ &= \sum ln[f(y_i|\theta)] &(6)\\ &= \ell(\theta|D) &(7)\end{aligned}$$

- $\ell(\theta|D)$ is the **log-likelihood function** of θ, given our data.
- Notice that the product in (5) turned into summation in (6). Computers are better able to deal with summation than multiplication, hence the logarithmic transformation.
- Importantly, the maximum of the log-likelihood function is the same as the maximum of the likelihood function. We lose no information!

Estimating π

We see 5 *heads* in 10 flips of a coin. What value of π produced this result?

- First we need to consider the distribution which gave rise to this data!
- Since we are considering flips of a coin, the data is generated from the *Binomial Distribution*:

$$f(heads, flips|\pi) = \mathcal{L}(\pi|heads, flips) =$$

$$= \frac{flips!}{heads!(flips-heads)!}\pi^{heads}(1-\pi)^{flips-heads}$$

- We thus need to find the maximum of the log-likelihood function: $ln[(10!/5!5!)\pi^5(1-\pi)^5]$ with respect to π.
- If you do the calculus, you should conclude that the maximum occurs when $\pi = 0.5$
- You have just derived your first ML estimator!

Intuition Behind Parameter Variance

We have now obtained a point estimate for π, but what about our certainty about the accuracy of this estimate?

- Logically, the steeper the (log-) likelihood function, the easier it is to find its maximum.
- The easier it is to declare the maximum, the more certain we are about this maximum.
- Thus, the larger the *curvature* of the (log-) likelihood function, the greater is our certainty of the estimate.
- Formally, this means that the larger the second partial derivative with respect to a given parameter, the more certain we are about the estimate of this parameter.
- In practice we tend to used the inverse of the negative expected values of the second partial derivatives to determine the variance-covariance matrix, but that is really a technical matter... (see Long p.32)

ML Estimation of Linear Regression

The model is: $y_i = \mathbf{x_i}\boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$

- We thus need to estimate the βs in the vector $\boldsymbol{\beta}$ (which includes a constant), as well as σ^2.
- Given the assumptions about the error distribution, the PDF of the dependent variable is: $f(y_i|\mathbf{x_i}, \boldsymbol{\beta}, \sigma^2) \sim N(\mathbf{x_i}\boldsymbol{\beta}, \sigma^2)$

  The **Normal PDF** is:
  $f(y_i|\mu, \sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}}) \exp[-\frac{1}{2\sigma^2}(y_i - \mu)^2]$

  The **likelihood function** is:
  $\mathcal{L} = (2\pi\sigma^2)^{-.5n} exp(-\frac{1}{2\sigma^2} \sum (y_i - \mathbf{x_i}\boldsymbol{\beta})^2)$

  The **log-likelihood function** is:
  $\ell(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{x}) = -.5nln(2\pi) - .5nln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x_i}\boldsymbol{\beta})^2$

- [See example in R]

Conclusion

- MLE is a powerful, theoretically rooted method.
- It allows estimation of many different classes of models, and thus is much more versatile than OLS.
- ML estimation centers around the (log-)likelihood function. This function is the basis for both point estimates, as well as confidence intervals and hypothesis testing.
- MLE is thus a unified method of statistical estimation.