CITY AND REGIONAL PLANNING 775 / CIVIL ENGINEERING 775

Understanding Maximum-Likelihood Estimation

Philip A. Viton

November 22, 2010

Contents

1 Introduction

This note tries to explain and motivate the fundamental idea behind maximum-likelihood estimation of parameters, via simple example.

2 Basic Idea

Suppose we have a random sample of size 2 (call it y_1 and y_2) from a $N(\mu, 1)$ distribution, that is, we know that the sample is drawn from a normal distribution with variance 1, but the mean is unknown. Our task is to estimate the mean.

Now, as soon as we choose a value for μ, we know the complete probability density of the data, and we can calculate the probability of observing our sample data, given that choice. We call this the *sample likelihood*. If $f(y)$ is the normal density or frequency function for a normal variate with mean μ and variance σ^2, that is,

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

and then, because we have a random sample, the sample likelihood is

$$L(\mu) = \Pr[\text{sample} \mid \mu] = f(y_1, \mu, 1) \cdot f(y_2, \mu, 1) \tag{1}$$

— that is, the sample likelihood is just the product of the individual probabilities.

Suppose we plug in a guess for μ, and we calculate the sample likelihood using equation (1). And suppose it turns out that the sample likelihood is very low. We now have two choices. We can conclude that our guess was right, but that something very unusual has happened. Or we can conclude that our initial guess was wrong, that is, that our guess was in a sense inconsistent with the data we have actually obtained. *The fundamental idea behind maximum likelihood estimation is the belief that the second conclusion is more reasonable.* It then makes sense to try to find a better guess, one that makes the observed data more "probable".

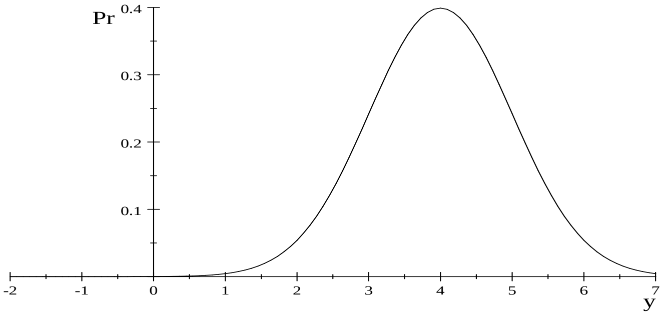
The *maximum likelihood estimator* of a parameter is that value of the unknown parameter that results in the probability of the observed data (the sample likelihood) being as high as possible. For computational reasons, it is often more convenient to think of maximizing the sample log-likelihood rather than the likelihood: since the logarithm is an increasing function, this will give the same answer as maximizing the likelihood itself.

3 Example

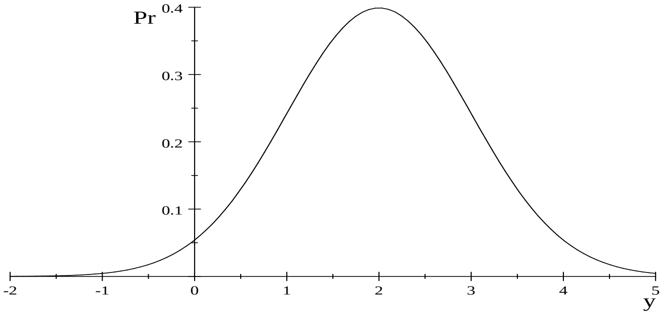
A numerical example may help make this clearer. Suppose our data is $y_1 = 0.5$ and $y_2 = 0.4$. Suppose we guess that the mean of the distribution is $\mu = 4$. A straightforward calculation (remember that we are assuming that $\sigma^2 = 1$) shows that

$$\begin{aligned} L(4) &= \Pr[y_1 = 0.5 \text{ and } y_2 = 0.4 \mid \mu = 4] = 5.33996 \times 10^{-7} \\ \log L(4) &= -14.4429 \end{aligned}$$

How are we to interpret this? Well, if we were drawing a random sample from the $N(4, 1)$ distribution, we *might* have got our observed data, $y_1 = 0.5$ and $y_2 = 0.4$. But as our calculation shows, this is not very likely, in fact, vanishingly probable. The plot below shows what is going on. The $N(4, 1)$ density is a bell-shaped curve centered on the estimated mean (here, 4). But our data lies in the region less than 1 on the horizontal axis, and as the picture shows, each of our sample points has essentially zero probability.



Adopting the maximum-likelihood principle, we therefore conclude that $\mu = 4$ was a bad guess. How could we do better? The picture suggests the answer: we should shift the curve to the left: this will result in our data having a higher probability. Since the normal curve is centered on its mean, this implies that we should try a lower guess for the mean of the distribution. For example, suppose we try $\mu = 2$. Then the picture is



and as we can see, the probability of our data points ($y_1 = 0.5$ and $y_2 = 0.4$) is now seriously higher. It also seems that we could do better still.

The table below shows the results of some initial guesses. Focussing on the log-likelihood column, we see that our successive guesses result in improvements (higher probability of observing our actual sample data) up until the last one, where the sample likelihood falls. This tells us that the maximum lies somewhere between the last three guesses.[1]

μ	**Pr**[$\mathbf{y_1}$]	**Pr**[$\mathbf{y_2}$]	**Sample Likelihood**	**Sample LogL**
4.0000	0.0006	0.0009	0.0000	−14.4429
2.0000	0.1109	0.1295	0.0144	−4.2429
1.0000	0.3332	0.3521	0.1173	−2.1429
0.5000	0.3970	0.3989	0.1584	−1.8429
0.3000	0.3970	0.3910	0.1552	−1.8629

The next table repeats the exercise, and tries to improve our estimate.

μ	**Pr**[$\mathbf{y_1}$]	**Pr**[$\mathbf{y_2}$]	**Sample Likelihood**	**Sample LogL**
0.8000	0.3683	0.3814	0.1405	−1.9629
0.6000	0.3910	0.3970	0.1552	−1.8629
0.4000	0.3989	0.3970	0.1584	−1.8429
0.3000	0.3970	0.3910	0.1552	−1.8629
0.4400	0.3986	0.3982	0.1587	−1.8405
0.4500	0.3984	0.3984	0.1588	−1.8404
0.4600	0.3982	0.3986	0.1587	−1.8405

As we can see, it looks very much as if the estimate $\mu = 0.45$ is our best guess, and is therefore the maximum-likelihood estimator for this problem (write this as $\hat{\mu}_{\text{MLE}}$). As it happens, this can be shown to be right: for this particular problem, it can be shown analytically[2] that the maximum-likelihood estimation of the population mean is the sample mean, which for our example is $\frac{0.4+0.5}{2} = 0.45$. But this won't help for our logit problem, where no analytical solution is possible, and we have to rely on our trial-and-error strategy, ie on numerical optimization.

[1] How do we know that this has really bracketed the maximum, and that the real answer doesn't lie somewhere else entirely? The answer is that in general we don't. These methods will yield an estimate only of a *local maximum*. Of course, given the shape of the normal distribution, it's pretty obvious that we will in fact find the correct answer; but in general we can't be sure. One strategy, if you have doubts, is to start from a variety of initial guesses. If the results all end up at roughly the same place, this can reinforce your confidence that you've found the true maximum — but you can never be 100% sure.

[2] This is an easy exercise in calculus: write down the sample likelihood in terms of μ (with $\sigma^2 = 1$), differentiate with repect to μ, set the result equal to zero, and solve for μ.The result is the MLE.

4 Theoretical considerations

We now have an intuitive justification for the maximum-likelihood estimate of a parameter: it's the value that makes it most likely that we get the data we actually observe. Can we make a stronger case? It turns out that we can. The first thing to note is that the maximum-likelihood estimator (MLE) of a parameter depends on the data points, which are random variables (they're a random sample). The MLE is therefore itself a random variable, and we can ask about its distributional properties. The important properties occur in the context of a large sample.[3] Specifically, we have the following:

1. As the sample size gets large the distribution of the MLE (its probability density or frequency function) looks like a very narrow spike centered on the true value of the parameter. Another way of saying this is that the probability of the MLE differing from the true value of the parameter approaches zero as the sample size increases. The technical term is that the MLE is a *consistent estimator* of the true parameter.

2. Though the MLE need not be unbiased, any bias disappears as the sample size gets large. We say that the MLE is *asymptotically unbiased*.

3. As the sample size get large, the shape of the distribution of the MLE is that of the normal distribution. We say that the MLE is *asymptotically normal.*

The first three properties are often summarized in the phrase that the MLE is CAN – consistent and asymptotically normal. But what about the variance of the distribution of the MLE? We have a further result, relating the MLE to other estimators of our unknown parameter:

4. In the context of large samples, and under some fairly general conditions that are usually satisfied in practical applications, the MLE has a lower variance than any other asymptotically unbiased estimator.[4] In this sense, the MLE is an optimal estimator.

[3] Technically, as the sample size approaches infinity. This raises the question of how large a real-world sample needs to be for these results to apply. There are no general results, but most people believe that a random sample of 100 or more is probably large in the sense needed here.

[4] For those who have had an advanced statistics course, this statement can be made a bit more precise. There is a theoretical lower bound on the variance of any unbiased estimator, known as the Craḿer-Rao lower bound. The MLE achieves this lower bound with equality.

Of course, for this to be practically useful, we need to estimate the variance of the MLE, but it turns out that this can be done, too.[5] Now suppose we obtain the MLE, $\hat{\mu}_{\text{MLE}}$, and estimate its variance say by $\hat{\sigma}^2_{\hat{\mu}}$. Consider the null hypothesis that the *true value* of the unknown parameter — the one we are estimating — is some given number r. Then under the null hypothesis, the statistic

$$\frac{\hat{\mu}_{\text{MLE}} - r}{\hat{\sigma}_{\hat{\mu}}}$$

has, in large samples, a student's t-distribution.[6] In other words, you can do hypothesis testing with the MLE just as you would do with the standard linear model.

So far we've discussed maximum-likelihood estimation in terms of a single parameter. But exactly the same considerations apply in the case where we want to estimate a vector of parameters. Of course, the task of finding the maximum likelihood is going to be more complicated; but there are computer methods that will take care of the messy details. The important thing is to understand what we're doing, and why.

5 ML and least-squares

If you think back to the case where we want to estimate the parameter vector of the linear model

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

you now seem to have two choices: the ordinary-least-squares estimator (OLS) and the MLE. But it turns out that when the error terms ε_i are iid $N(0, \sigma^2)$ random variates, the OLS estimator and the MLE of β are the same.[7] However, this is not true in general.[8]

[5] You can do this via the second derivative of the log-likelihood function, or, in the case of many parameters, of the matrix of second derivatives, known as the Hessian matrix.

[6] The degrees of freedom are [sample size] -1. It is worth remembering, for practical applications, that when the degrees of freedom exceeds about 30, the t-distribution approaches the standard normal distribution. (That is why you rarely see tables of the t distribution with many degrees of freedom).

Useful rule-of-thumb: if the computed value of the test statistic is greater than 2 in absolute value, we *reject* the null hypothesis at the 5% signficance level.

[7] You may recall that the OLS estimator also has an optimality property: it is best in the class of *linear* unbiased estimators. The MLE relaxes the requirement of linearity, but needs a large sample for its optimality property.

[8] In fact, the OLS estimator and the MLE of the *variance* σ^2 are different, and the MLE of is biased, while the OLS estimator is not.

6 ML for logit models

It is easy to write down the likelihood of our data for a random sample of discrete choices under the logit model. If

$$y_{ij} = \begin{cases} 1 & \text{if individual } i \text{ chose mode } j \\ 0 & \text{otherwise} \end{cases}$$

is an indicator of the choices actually made by our sample, then the sample likelihood is[9]

$$L = \prod_{i=1}^{I} \prod_{j=1}^{J} {P_{ij}}^{y_{ij}}.$$

In the case of the logit model we assume that the systematic part of utility is linear-in-parameters, so the probability that individual i selects mode j is:

$$P_{ij}(\beta) = \frac{e^{x_{ij}\beta}}{\sum_k e^{x_{ik}\beta}}.$$

Then, recalling that logarithms convert products to sums, the sample log-likelihood is

$$\log L(\beta) = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} \ln(P_{ij}(\beta))$$

which is a function of the vector β, and all we have to do is find values for β that maximize this expression. It turns out that this is a well-behaved problem, and is very easy to do numerically, even for large samples (many individuals) facing many alternatives. For all practical purposes, we just need to feed our data to a compute program, press a key, and we're almost instantly rewarded with our answer.

[9] This should give you a clue as to how to relax the assumption that all people in our sample face the same choice set. The basic idea is to define an individual-i-specific choice set of dimension J_i. Though the notation is now a bit more complicated, nothing essential changes.