Statistics 3858 : Maximum Likelihood Estimators

1 Method of Maximum Likelihood

In this method we construct the so called likelihood function, that is

$$L(\theta) = L(\theta; X_1, X_2, \ldots, X_n) = f_n(X_1, X_2, \ldots, X_n; \theta)$$

The function f_n is either the joint pdf or joint pmf of the random variables $X_1, X_2, \ldots, X_m$, and the notation denotes the dependence of this distribution on θ, where $\theta \in \Theta$.

The notation for the likelihood function L usually suppresses the random data $X_1, X_2, \ldots, X_n$ and we usually write this as $L(\theta)$. Sometimes we also wish to denote the dependence on n and write $L_n(\theta)$. When we want to denote the dependence on the random variables $X_1, X_2, \ldots, X_n$ we write $L(\theta; X_1, X_2, \ldots, X_n) = f_n(X_1, X_2, \ldots, X_n; \theta)$.

When we have observed data $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, where the lower case letters denote the specific observed data we then have the observed likelihood

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = f_n(x_1, x_2, \ldots, x_n; \theta)$$

For given random variables the maximum likelihood estimator, say $\hat{\theta}_n$ is the argument $\theta \in \Theta$ for which

$$L(\hat{\theta}_n) = \max_{\theta \in \Theta} L(\theta)$$

Notice this means that

$$\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$$

which says that $\hat{\theta}_n$ is a function of the random data or random variables $X_1, X_2, \ldots, X_n$. Sometimes we may find this explicitly by finding this function, say h so that

$$\hat{\theta}_n = h(X_1, X_2, \ldots, X_n)$$

For observed data $x_1, x_2, \ldots, x_n$ this says

$$\hat{\theta}_n = h(x_1, x_2, \ldots, x_n)$$

Mathematically this says we maximize over possible θ the function

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = f_n(x_1, x_2, \ldots, x_n; \theta)$$

by treating this is a function with argument θ and treating $x_1, x_2, \ldots, x_n$ as given and known numbers.

The method of moments searches for a value of the parameter θ that matches the sample moments. The method of maximum likelihood searches for the value of the parameter θ that gives the largest probability (or density) for the observed data.

Since our goal is to maximize the likelihood can use anything proportional to the joint pdf or joint pmf, provided the constant of proportionality does not involve θ. Therefore we use as the likelihood function L anything positively proportional to the joint pdf or joint pmf

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) \propto f_n(x_1, x_2, \ldots, x_n; \theta) \tag{1}$$

The notation for (1) emphasizes that we are dealing with a function with argument $\theta \in \Theta$, and the second expression is used when we need to emphasize that it also depends on the observed data (or the observable r.v.s with X in place of x).

Let $\hat{\theta}_n$ be the element $\theta \in \Theta$ which maximizes $L(\theta) = L(\theta; X_1, \ldots X_n)$. We denote this value as

$$\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} L(\theta) \ .$$

Sometimes we are interested in a particular subset of Θ and the maximization may be done over that set. In any case the idea is the same but the maximization is restricted to θ in this particular subset. For example in a normal model the parameter space is $\Theta = R \times R^+$, but we may require the mean is 0 and so compute a restricted MLE over the set $\Theta' = \{(0, \sigma^2) : \sigma^2 > 0\} \subset \Theta$. Of course we could also work a new statistical model with parameter space R^+, where the parameter corresponds to σ^2.

It is again worth pointing out that at this stage of finding the argument that maximizes the likelihood, we find this formula by either working with $L(\theta, X_1, \ldots, X_n)$ or with $L(\theta, x_1, \ldots, x_n)$, as either will give the formula to produce the function that gives $\hat{\theta}$. One is the r.v.s $\hat{\theta}(X_1, \ldots, X)$ and the other is simply the function $\hat{\theta}(x_1, \ldots, x_n)$. Of course when we plug in the observed data both will give the same value, as they are based on the same function. We had the same issue when working with the method of moments.

The most common setting is when the random variables are independent, and more specifically are iid. In this case if the marginal distribution of X_i is $f(\cdot; \theta)$ (either pdf of pmf) where θ is some value in the parameter space, then the likelihood function is

$$L(\theta) = L(\theta; X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(X_i; \theta) \ .$$

This is because in the case that X_i are iid with distribution $f(\cdot; \theta)$ the joint distribution of $X_1, X_2, \ldots, X_n$ is given by the function (with arguments $x_1, x_2, \ldots, x_n$)

$$f_n(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \ .$$

If the X_i's are dependent r.v.s, for example an AR(1) sequence of r.v.s or another Markov process, then the joint distribution f_n will be of a different form.

2 Some Examples

Binomial

For data $X \sim \text{Binom}(n, \theta)$ with $\theta \in \Theta = [0, 1]$ the pmf is

$$f(x; \theta) = \binom{n}{x}(1-\theta)^{n-x}\theta^x$$

for $x = 0, 1, \ldots, n$. The likelihood is then taken to be either

$$L(\theta) = \binom{n}{X}(1-\theta)^{n-X}\theta^X$$

or

$$L(\theta) = (1-\theta)^{n-X}\theta^X$$

How do we find $\hat{\theta}_n$? L is a function of a single argument θ, it is differentiable, so we may use calculus methods to find the MLE. For the purpose of finding this formula is does not matter whether we use X as the random variable (or random vector as needed) or x as the observed data (usually vector of data) x_{obs} Our object in finding the formula for $\hat{\theta}_n$ is simply to find this formula, which will involve X or x. When we wish to study $\hat{\theta}_n$ as a r.v. we need to use it in the form $\hat{\theta}_n(X)$ and when we wish to calculate the observed value, that is the *estimate*, we use it in the form $\hat{\theta}_n(x)$. Once we find the formula $h(x)$ we then use it to obtain the estimate $h(x_{\text{obs}})$ or the r.v. $h(X)$.

$$\begin{aligned}
\frac{dL(\theta)}{d\theta} &= \frac{d\{(1-\theta)^{n-x}\theta^x\}}{d\theta} \\
&= (n-x)(1-\theta)^{n-x-1}(-1)\theta^x + x(1-\theta)^{n-x}\theta^{x-1} \\
&= \{-(n-x)\theta + x(1-\theta)\}(1-\theta)^{n-x-1}\theta^{x-1}
\end{aligned}$$

Recall from calculus that the maximum value occurs either at $\hat{\theta}$ which (i) is the solution of

$$\frac{dL(\theta)}{d\theta} = 0 \tag{2}$$

or (ii) $\hat{\theta}$ is on the boundary of Θ, that is 0 or 1 in this case.

The student should review this maximization method from their calculus text or notes for their calculus of one variable and calculus of several variables.

Consider solving (2). There are actually 3 solutions, say

$$r_1 = 0, r_2 = 1, r_3 = \frac{x}{n} \ .$$

Notice that if $x = 0$ then $r_3 = 0$ and if $x = n$ then $r_3 = 1$.

Consider the cases

- $x = 0$. The student should sketch the function $L(\theta)$ and verify there is an unique maximizer, $\hat{\theta}_n = 0$. You may also see the remark after this example. In this case

  $$L(\theta) = (1-\theta)^n \ .$$

  Is there any θ which makes the derivative zero?

- $x = n$. The student should sketch the function $L(\theta)$ and verify there is an unique maximizer, $\hat{\theta}_n = 1$.
- $0 < x < n$. In this case $L(0) = 0$, $L(1) = 0$ and $L(\frac{x}{n}) > 0$, so that argmax $= \frac{x}{n}$.

Fortunately in all three cases we have

$$\text{argmax}_{\theta\in\Theta} L(\theta) = \hat{\theta}_n = \frac{x}{n}$$

so we obtain the one algebraic solution.

Continuation

Since differentiating a sum is easier than differentiating products, we may use a strictly monotone increasing transformation of $L(\theta)$ and maximize it instead. The argmax of this will be the same as the argmax of L. Since log of a product is a sum of the logs we use a logarithmic transformation, specifically the natural log, so that differentiation is easier. The so called log likelihood is then

$$\Lambda(\theta) = \log(L(\theta)) = (n - X)\log(1-\theta) + X\log(\theta) + \log\left(\binom{n}{X}\right)$$

In this form $\Lambda(X)$ is a random function, so the solution $\hat{\theta} = \text{argmax}(\theta)\{\Lambda(\theta)\}$ is a random variable.

In order to find $\text{argmax}(\theta)\{\Lambda(\theta)\}$ we can equally well work with

$$\Lambda(\theta) = \log(L(\theta)) = (n - x)\log(1-\theta) + x\log(\theta) + \log\left(\binom{n}{x}\right)$$

where instead of the r.v. X we use a generic value x or the appropriate type, in this case an integer from $0, \ldots, n$. Either of these forms allows us to find the formula for $\hat{\theta} = \text{argmax}(\theta)\{\Lambda(\theta)\}$. When we want to study this as a r.v. we must work with $\hat{\theta}(X)$, and when we just want to specify the function we can work with $\hat{\theta}(x)$. Notice these two, while looking similar, one is the r.v. that is a function of X whereas the other is simply a function. The same notion applies to $L(\theta, X)$ and $L(\theta, x)$, where we explicitly put in the dependence on the r.v. X or the (generic) argument x.

Now return to the log likelihood. The derivative wrt θ (when $x \neq 0$ or $x \neq n$) is

$$\begin{aligned}\frac{d\log(L(\theta))}{d\theta} &= \frac{d\{(n-x)\log(1-\theta) + x\log(\theta)\}}{d\theta} \\ &= -\frac{n-x}{1-\theta} + \frac{x}{\theta}\end{aligned}$$

$$\frac{d\log(L(\theta))}{d\theta} = 0$$

yields

$$\hat{\theta}_n = \frac{x}{n}\ .$$

Fortunately this formula does give the correct maximizer even in the cases $x = 0, n$.

Notice that we have the argmax is of the form $h(x)$, that is a function of the data. To study this as a r.v. we study $h(X)$ and to obtain the *estimate* we use $h(x)$.

Also notice that if we had used

$$\Lambda(\theta) = \log(L(\theta)) = (n - X)\log(1-\theta) + X\log(\theta)$$

we would get the same solution for argmax, that is we omit the term $\log\left(\binom{n}{x}\right)$ which is constant with respect to θ.

The maximum likelihood estimator is

$$\hat{\theta}_n = \frac{X}{n} = \bar{X}_n$$

In this statistical model the MLE is the same as the method of moments estimator.

End of Example

Remark

In the above example we use log likelihood since the algebra is easier. We do need to be careful how we use calculus. In particular if x is not 0 or n then

$$L(0) = L(1) = 0$$

and so we cannot take logarithms when θ equals 0 or 1.

If $x = 0$ then

$$\begin{aligned} L(\theta) &= (1-\theta)^{n-x}\theta^{x} \\ &= (1-\theta)^{n-0}\theta^{0} \\ &= (1-\theta)^{n} \end{aligned}$$

In particular $L(1) = 0$. Taking logs we then have, except for $\theta = 1$

$$\log(L(\theta)) = n\log(1-\theta)$$

Differentiating gives

$$\frac{d\log(L(\theta))}{d\theta} = -\frac{n}{1-\theta}$$

and there is no solution to

$$\frac{d\log(L(\theta))}{d\theta} = -\frac{n}{1-\theta} = 0$$

According to the calculus method of finding maximizers then argmax$(\log(L(\theta))$ must occur at either 0 or 1. Since $\log(L(1)) = \lim_{\theta\to 1^-}\log(L(\theta)) = -\infty$ and $\log(L(0)) = \log(1) = 0$ then argmax equals 0.

Before taking logs we could also have noted when $x = 0$ that $L(\cdot)$ is a strictly monotone decreasing function of argument θ and so the maximum has to occur at 0.

Fortunately often the formal calculus does still give the correct answer, but not always. Later we will have to consider a so called regular maximum likelihood case and the non-regular case. The binomial example will be a regular case as we can obtain the MLE using calculus methods since since the two special cases of $x = 0$ and $x = n$ have a solution of the same form as all other $0 < x < n$.

End of Remark

Remark Why do we use natural log for our log likelihood? Here we use log for natural log. Why not log base 10? For $x > 0$

$$x = e^{\log(x)} = 10^{\log_{10}(x)} = \left(e^{\log(10)}\right)^{\log_{10}(x)} = e^{\log(10)\log_{10}(x)}$$

Thus

$$\log(10)\log_{10}(x) = \log(x)$$

Therefore

$$\frac{d\log_{10}(x)}{dx} = \frac{1}{\log(10)}\frac{d\log(x)}{dx} = \frac{1}{x\log(10)}$$

Differentiating natural logs is algebraically more convenient than differentiating logs of another base.

End of Remark

Normal Example

For convenience we write either $\theta = (\mu, \sigma^2) = (\theta_1, \theta_2)$ in order to emphasize one aspect of our calculation below. In this formulation we treat the parameter as (mean , variance). That means that derivatives are with respect to (μ, σ^2) and not with respect to (μ, σ).

$$L((\theta_1, \theta_2)) = \frac{1}{\theta_2^{\frac{n}{2}}} e^{-\frac{1}{2\theta_2}\sum_{i=1}^{n}(X_i - \theta_1)^2}$$

and the log likelihood is

$$\log(L((\theta_1, \theta_2)) = -\frac{n}{2}\log(\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(X_i - \theta_1)^2$$

Again we note that for the next part, where we find the MLE it does not matter whether we replace X_i with x_i as we are just finding the formula for the MLE. It only makes a difference when we either wish to study $\hat{\theta}_n$ as a random variable or as a estimate, that is when we need to distinguish the difference between estimator and estimate.

$$\begin{aligned}\frac{\partial \log(L(\theta_1, \theta_2))}{\partial \theta_1} &= \frac{1}{\theta_2}\sum_{i=1}^{n}(X_i - \theta_1)\\ \frac{\partial \log(L(\theta_1, \theta_2))}{\partial \theta_2} &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i - \theta_1)^2\end{aligned}$$

Setting

$$\begin{aligned}\frac{\partial \log(L(\theta_1, \theta_2))}{\partial \theta_1} &= 0\\ \frac{\partial \log(L(\theta_1, \theta_2))}{\partial \theta_2} &= 0\end{aligned}$$

and solving gives

$$\begin{aligned}\hat{\theta}_{1n} &= \bar{X}_n\\ \hat{\theta}_{2n} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\end{aligned}$$

Reverting to our usual parameter notation $\theta_1, \theta_2) = (\mu, \sigma^2)$ we have

$$\begin{aligned}\hat{\mu}_n &= \bar{X}_n\\ \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\end{aligned}$$

Notice $\hat{\sigma}_n^2$ is not equal to the sample variance, and it is a biased estimator of σ^2.

Continuation : What would happen if instead we used the parameterization for the normal family of (mean, standard deviation)?

In that case we would be differentiating w.r.t. (μ, σ), but also our log likelihood would be

$$\log(L((\mu,\sigma)) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \ .$$

The student should work through this parameterization and find that

$$\begin{aligned} \hat{\mu}_n &= \bar{X}_n \\ \hat{\sigma}_n &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2} \ . \end{aligned}$$

Aside : There is a property of likelihood estimation that as long as one reparameterizes by a 1 to 1 continuous form the MLE are equivalent for either form. Here above we see this is the case.

End of Example

In our calculations for MLE we will no longer concern ourselves with different parameterizations, and just work with the one that is natural for our model. The point of this comment is that as long as we work consistently it will not matter which possible parameterization we use.

Uniform($0, \theta$*)* , $\theta \in \Theta = (0, \infty)$

Here we consider an example this is not *regular*. It is not possible to find the MLE using the calculus tool of differentiation to find local extremes.

Here we consider $X_i, i = 1, 2, \ldots, n$ iid from a Uniform$(0, \theta)$ distribution and $\theta \in \Theta = (0, \infty)$. Our goal is to find the MLE.

The student should verify that the method of moments estimator is

$$\tilde{\theta}_n = 2\bar{X}_n$$

and that it is an unbiased estimator of θ. Also find the MSE of $\tilde{\theta}$.

To motivate this consider first the case $n = 1$. It is not interesting in itself, but it will help us to understand the likelihood function. Here

$$L(\theta) = f(x;\theta) = \frac{1}{\theta}\mathrm{I}(0 < x \le \theta)$$

Note it does not matter if we include of exclude the endpoints in this function with argument x, but we include it to simply some algebra.

In the pdf we treat x as the argument for a given parameter θ. In the likelihood we treat θ as the argument for a given x. For $\theta = 2$ the student should sketch the pdf. For $x_{\text{obs}} = 2$ the student should sketch the likelihood $L(\theta) = L(\theta; x = 2)$. These are also plotted in Figures 1 and 2. Notice that $L(\theta) = 0$ if $\theta < x_{\text{obs}}$ and $L(\theta) = \frac{1}{\theta}$ if $\theta \ge x_{\text{obs}}$.

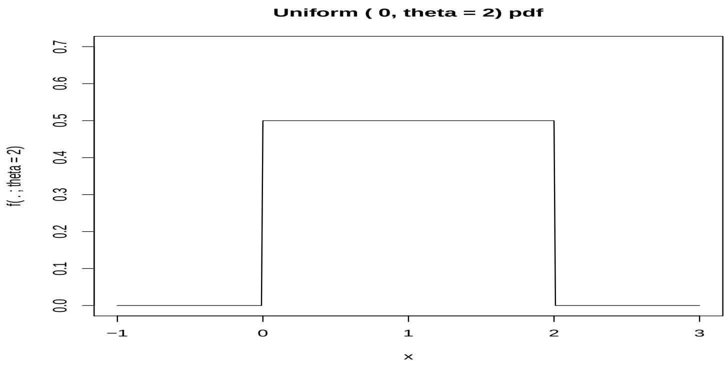

Figure 1: Uniform$(0, \theta = 2)$ pdf

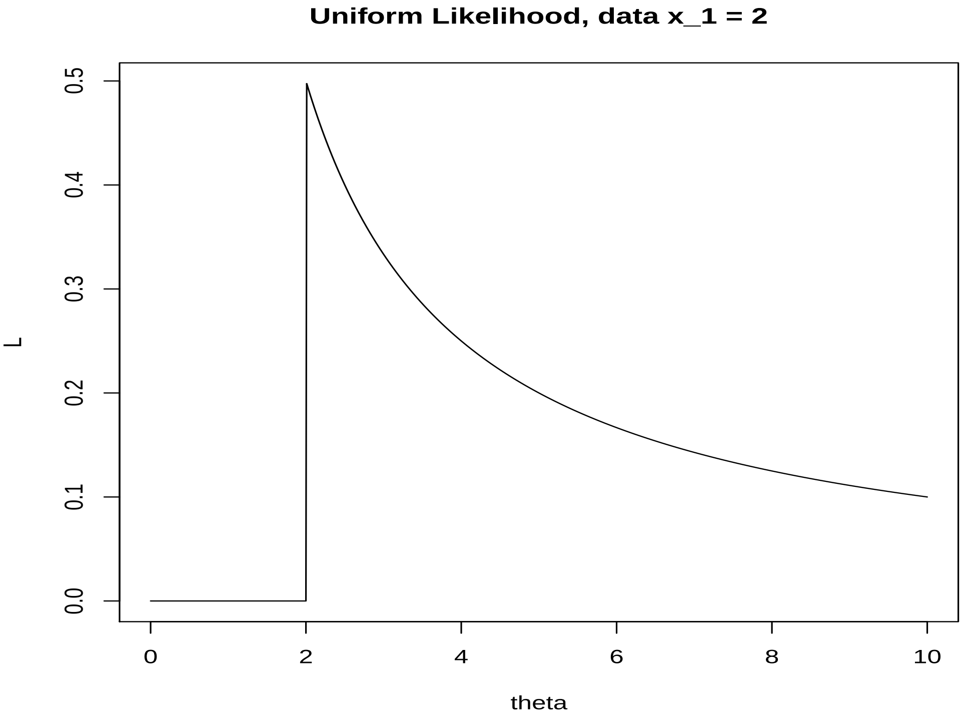

Figure 2: $L(\theta; x = 2)$

For $n > 1$ we have the a similar property. $L(\theta)$ evaluates to 0 for some θ, but which ones?

$$\begin{aligned} L(\theta) &= L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \left\{ \frac{1}{\theta} \mathrm{I}(0 < x_i \leq \theta) \right\} \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} \{\mathrm{I}(0 < x_i < \theta)\} \end{aligned}$$

Can we simplify this expression, that is the product in the last line? This product is either 1 or 0. It is one only if all the indicators evaluate to 1. This means that for every i we must have $x_i \leq \theta$, or equivalently $\theta \geq \max\{x_1, \ldots, x_n\} = x_{(n)}$, the maximum order statistic. The likelihood function is (using $x_i > 0$)

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n} \mathrm{I}(x_{(n)} \leq \theta)$$

In this statistical model the MLE is $\hat{\theta}_n = X_{(n)}$, the maximum order statistic, and the method of moments estimator is $\tilde{\theta}_n = 2\bar{X}_n$. We see these two estimators are different.

Notice also that the likelihood function is not

$$L(\theta) = \frac{1}{\theta^n}$$

as this expression is not complete. Suppose we did use this *incorrect* expression and tried to maximize this using the usual calculus tool. Taking logs and differentiating we need to solve

$$0 = \frac{d\{n \log(\theta)\}}{d\theta} = \frac{n}{\theta}$$

There is no solution. This does not mean that we cannot maximize $L(\theta)$ or $\log(L(\theta))$. The calculus method of determining the maximum of the function $L(\theta)$ does not apply. The student should review the calculus tool for maximizing functions.

At the beginning of this example it is stated this example is not *regular*. This term will be explained later, but for practical purposes *regular* models will allow us to use calculus methods to find argmax, and non-*regular* models require a non-calculus method to find argmax.

End of Example

Normal Regression Model

Our model is of the form, for given covariates $x_1, \ldots, x_n$ the r.v.s's Y_i are independent and satisfy

$$Y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i$$

where ϵ_i are iid $N(0, \sigma^2)$. The data are of the form of pairs (x_i, Y_i) where the x_i are specified numbers (specified by the experimenter). The joint distribution of $Y_1, \ldots, Y_n$ for the given covariates $x_1, \ldots, x_n$ is then determined by the n different marginal distributions

$$f_{Y_i}(y; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x_i)^2} \ .$$

Since the ϵ_i are iid then Y_i are independent, but they are not identically distributed, since they have different means. We treat the parameter as $\theta = (\beta_0, \beta_1, \sigma^2)$ since these are the unknowns in the model that we want to estimate, while the $x_1, \dots, x_n$ are a different type of object, similar to parameters but are known or determined by the experimenter. The parameter space is $\Theta = R \times R \times R^+$.

The log likelihood is

$$\log(L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

The MLE is the solution in terms of $(\beta_0, \beta_1, \sigma^2)$ of

$$\begin{aligned} 0 = \frac{\partial \log(L(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \\ 0 = \frac{\partial \log(L(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) \\ 0 = \frac{\partial \log(L(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

The student should finish this calculation to obtain the MLE.

End of Example

Gamma Model

Consider X_i, $i = 1, \dots, n$ iid from a Gamma pdf, that is X_i has pdf

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} \mathrm{I}(x > 0)$$

where $\theta = (\alpha, \lambda) \in \Theta = R^+ \times R^+$.

The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{\lambda^\alpha X_i^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda X_i}$$

Since $X_i > 0$ with probability 1 for every θ, we can calculate the log likelihood for any θ, and now write $\theta = (\alpha, \lambda)$

$$\log(L(\alpha, \lambda)) \quad = \quad n\alpha \log(\lambda) + (\alpha - 1) \sum_{i=1}^{n} \log(X_i) - n \log(\Gamma(\alpha)) - \lambda \sum_{i=1}^{n} X_i$$

Recall the function Γ is given through the formula

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

Except for integer values of α there is no simple closed for expression to help us evaluate the Gamma function. However one can show the Gamma function is differentiable with respect to its argument α, and in fact has derivatives of higher order, so that in particular it has first and second derivatives wrt α.

Since X_i cannot be 0 or ∞ the argmax of the loglikelihood cannot occur on the boundary of Θ and therefore the MLE satisfies

$$0 = \frac{\partial \log(L(\alpha, \lambda)}{\partial \alpha} \quad = \quad n \log(\lambda) + \sum_{i=1}^{n} \log(X_i) - n \frac{d \log(\Gamma(\alpha))}{d\alpha}$$

$$0 = \frac{\partial \log(L(\alpha, \lambda)}{\partial \lambda} \quad = \quad \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i$$

We see that

$$\hat{\lambda}_n = \frac{\hat{\alpha}_n}{\bar{X}_n}$$

but of course we still need to find $\hat{\alpha}_n$.

One thing that we can see is that the MLE is not the same as the method of moments estimator.

For given observed data $x_1, \ldots, x_n$ the MLE is calculated by a numerical method. The most common method is the Newton-Raphson method, an iteration method. Computer programs or code are needed to evaluate these MLE. Many programs and languages implement these, such as R, S, Mathlab as well as many others. Sometimes stand alone code is available for programming in lower level languages such as FORTRAN, C or C+.

End of Example

Remark

The functions

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} e^{-x} dx \\ \text{digamma}(\alpha) &= \frac{d \log(\Gamma(\alpha))}{d\alpha} \\ \text{trigamma}(\alpha) &= \frac{d^2 \log(\Gamma(\alpha))}{d\alpha^2} \end{aligned}$$

are useful in solving for the MLE, but these are also not algebraic, that is there is not any simple closed for expression for these. The functions are evaluated by numerical methods in computer code. In particular these are implemented in many computer languages, such as R, S and Matlab.

End of Remark

Multinomial Model

Recall a multinomial random vector is the vector of counts in M catagories, say $S = \{1, 2, \ldots, M\}$, for n iid trials, and $p = (p_1, p_2, \ldots, p_M)$ is the probability distribution on the M categories so that for an individual trial p_j is the probability that the trial produces outcome in category j. This give a random vector $N = (N_1, N_2 \ldots, N_M)$ with a multinomial distribution with parameter p.

The parameter $p \in \Theta$, the set

$$\Theta = \{p = (p_1, p_2, \ldots, p_M) : p_j \geq 0\ ,\ p_1 + p_2 + \ldots + p_M = 1\}$$

This set is often called the simplex of order M. It is a set with dimension $M - 1$. This is because there are $M - 1$ degrees of freedom, in the sense that one must specify $M - 1$ numbers and the last one is then determined, that is the last number has no freedom as to possible values.

For integers $n_1, \ldots, n_M$ such that $n_j \geq 0$ and $n_1 + \ldots + n_M = n$

$$P\left(N = (n_1, n_2 \ldots, n_M); p\right) = \frac{n!}{n_1! n_2! \ldots n_M!} p_1^{n_1} p_2^{n_2} \ldots p_M^{n_M}$$

For observed data $N = (n_1, n_2 \ldots, n_M)$ the likelihood is

$$L(p) = p_1^{n_1} p_2^{n_2} \ldots p_M^{n_M}$$

and the log likelihood is

$$\log(L(p)) = \sum_{j=1}^{M} n_j \log(p_j)$$

To calculate the MLE we need to maximize $\log(L(p))$ over the parameter space Θ. Notice the M parameters p_j are not all algebraically free, so we must perform a constrained maximization.

The student should return to their advanced calculus notes or text and review the method of Lagrange multipliers.

Lagrange multipliers is a very useful tool for maximization with equality of inequality constraints. It involves making a new objective function by taking the original objective function and adding new Lagrange parameters times the equality (or inequality) constraints.

In our problem we make the new objective function

$$g(p_1, \ldots, p_M, \lambda) = \log(L(p)) + \lambda \left(\sum_{\ell=1}^{M} p_\ell - 1 \right) = \sum_{\ell=1}^{M} n_\ell \log(p_\ell) + \lambda \left(\sum_{\ell=1}^{M} p_\ell - 1 \right)$$

It is a function with $M + 1$ arguments, which we treat as algebraically free arguments. We maximize this using calculus tools. For this Lagrange problem we of course have the *solution* $\hat{p}_1, \ldots, \hat{p}_M, \hat{\lambda}$. Notice that $\hat{p}_j$ is then also an explicit function of $\hat{\lambda}$, say $\hat{p}_j(\hat{\lambda})$. In particular the $\hat{p}_j(\hat{\lambda})$ satisfies the original constraint. Langrange's Theorem concludes that $(\hat{p}_1(\hat{\lambda}), \ldots, \hat{p}_M(\hat{\lambda}))$ solves for argmax in the original constrained maximization problem.

$$\begin{aligned} \frac{\partial g}{\partial p_j} &= \frac{n_j}{p_j} + \lambda \\ \frac{\partial g}{\partial \lambda} &= \sum_{ell=1}^{M} p_\ell - 1 \end{aligned}$$

Multiplying the first by p_j and then summing these over $j = 1, \ldots, M$ gives

$$0 = \sum_{j=1}^{M} (n_j + \lambda p_j) = n + \lambda \sum_{j=1}^{M} p_j = n + \lambda$$

and where in the last part we use

$$0 = \frac{\partial g}{\partial \lambda} = \sum_{ell=1}^{M} p_\ell - 1 \ .$$

Therefore $\lambda = -n$ and hence

$$0 = \frac{n_j}{p_j} + \lambda = \frac{n_j}{p_j} - n$$

Rearranging this last expression gives the solution for p_j as $\hat{p}_{j,n} = \frac{n_j}{n}$.

Thus the maximum likelihood estimate is

$$\hat{p}_{j,n} = \frac{n_j}{n}$$

and the maximum likelihood estimator is

$$\hat{p}_{j,n} = \frac{N_j}{n}$$

End of Example

Aside

You can solve the above maximization without Lagrange multipliers by working with arguments $p_1, \ldots, p_{M-1}$ and the substituting $p_M = 1 - (p_1 + \ldots + p_{M-1})$. You then treat the log likelihood as a function of $M - 1$ arguments and proceed with the maximization as usual. Of course you pay for this with the complication of the care needed to calculate

$$\frac{\partial\ p_M(p_1, \ldots, p_{M-1})}{\partial p_j} = \frac{\partial(1 - p_1 - \ldots - p_{M-1})}{\partial p_j} = -1$$

for each $j = 1, \ldots, M - 1$. Much more care has to be taken in differentiating the log likelihood.

End of Aside

Multinomial $M = 2$

In this case for the random vector $N = (N_1, N_2)$, the r.v. N_2 equals the number of observations in category 2. If the categories are 1 = failure and 2 = success, then N_2 is the number of success in n iid trials with outcomes failure or success. N_2 then has a Binomial(n, p_2) distribution. The r.v. $N_1 = n - N_2$. In terms of the MLE we either

- maximize using Lagrange multipliers

$$g(p_1, p_2, \lambda) = n_1 \log(p_1) + n_2 \log(p_2) + \lambda(p_1 + p_2 - 1)$$

- using $\theta = p_2$, so that $p_1 = 1 - \theta$, and note that $n_1 = n - n_2$, and the maximize

$$L(\theta) \equiv L(1 - \theta, \theta) = (n - n_2) \log(1 - \theta) + n_2 \log(\theta)$$

Except for the change in notation this is the same as log likelihood for the binomial model.

End of Example

Multinomial with Further Constraints

Sometimes one is interested in a multinomial model in which $p = (p_1, \ldots, p_M)$ is further restricted than being a general member of the simplex of order M. For example it might be that (for $M = 3$) that

$$(p_1, p_2, p_3) = \left((1 - \theta)^2, 2\theta(1 - \theta), \theta^2\right) \equiv p(\theta) = (p_1(\theta), p_2(\theta), p_3(\theta))$$

where $\theta \in [0, 1]$. It is worth noting that this also is the Binomial$(2, \theta)$ probability vector with $0 \leq \theta \leq 1$.

Continuing with multinomial example notation we then find the log likelihood is

$$L(\theta) = \sum_{j=1}^{M} n_j \log(p_j(\theta)) = 2n_1 \log(1 - \theta) + 2n_2 \log(\theta(1 - \theta)) + 2n_3 \log(\theta)\ .$$

Notice that this log likelihood maximization is an unconstrained maximization wrt $\theta \in \Theta = [0,1]$. The student should complete this and find that

$$\hat{\theta}_n = \frac{2n_3 + n_2}{2n} \; .$$

End of Example

See problem 8.10.55 and 8.10.56 for another example of this type of constrained multinomial.

3 Beginning of Distribution Theory and Applications

The general order of useful calculations for the distribution of an estimator $\hat{\theta}_n$ are

1. Exact distribution - analytic or algebraic determination of the distribtution
2. Approximate methods, usually giving a normal approximation. This means that one works with either a centred or standardized distribution. Sometimes the standardization uses the population variance and sometimes with an estimator or estimate of the population variance. That is one typically works with
$$\sqrt{n}(\hat{\theta}_n - \theta)$$
or
$$\sqrt{n}\frac{(\hat{\theta}_n - \theta)}{v_n}$$
where v_n is an estimator (or estimate) of a population variance or $\frac{v_n^2}{n}$ is the estimated standard error of the estimator. This also typically requires knowing that the sample size n is large enough so that the normal approximation is reasonable, in the sense it can be used to give appropriate quantiles or critical value as needed, for example for confidence intervals.
3. An iterative calculation to obtain the estimate for given data combined with an approximation, again usually a normal approximation of a form as in 2 above
4. A parametric bootstrap approximation to the distribution of one of (i) $\hat{\theta}_n$ or (ii) a form as in 2 above.

One very important special case of the first type is when one observes iid data from a normal distribution. Recall the MLE is this case is

$$\begin{aligned} \hat{\mu}_n &= \bar{X}_n \\ \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \end{aligned}$$

We also have some Theorems from Chapter 6 (the student should review Section 6.3, including conditions for these theorems) so that in this case of iid $N(\mu, \sigma^2)$ r.v.s

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1)$$

and

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2_{(n-1)} \tag{3}$$

and these two r.v.s are independent. Therefore

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$$

and

$$\frac{n\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Also (Corollary B p 198)

$$t = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t_{(n-1)} \tag{4}$$

where S_n^2 is the sample variance. Notice also that

$$\hat{\sigma}_n^2 = \frac{n-1}{n} S_n^2 \ .$$

Formula (3) and (4) have some important implications in terms of obtaining a confidence interval. This is because these distributions are independent of the parameter (μ, σ^2) in the normal family. This property is a *pivotal* property and the random variables with this property are called *pivotal* quantities.

The formula for a 100(1 - α)% confidence interval for μ is obtained as the set of μ satisfying

$$t_{\frac{\alpha}{2},(n-1)} \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \leq t_{1-\frac{\alpha}{2},(n-1)}$$

Since the t distribution for any degree of freedom is symmetric about 0, then

$$t_{\frac{\alpha}{2},(n-1)} = -t_{1-\frac{\alpha}{2},(n-1)}$$

and the CI is

$$\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2},(n-1)} \leq \mu \leq \bar{X} + \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2},(n-1)}$$

Similarly a 100(1 - α)% confidence interval for σ^2 is obtained as the set of σ^2 satisfying

$$\chi^2_{\frac{\alpha}{2},(n-1)} \leq \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2},(n-1)}$$

This may be rewritten in terms of the sample variance as

$$\chi^2_{\frac{\alpha}{2},(n-1)} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2},(n-1)}$$

Solving for σ^2 yields

$$\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2},(n-1)}}$$