Maximum Likelihood

The method of Maximum Likelihood.

In developing the least squares estimator - no mention of probabilities.
Minimize the distance between the predicted linear regression and the observed data.
Need

- assumed normality or
- appeal to large sample results

to have results about distributions of the OLS estimator.
Maximum Likelihood: start in the opposite end.
Make probability assumptions: Assume we know the probability distribution.
Then find parameters that make the observed data "most likely" to have been observed.

Maximum Likelihood - evaluation relative to OLS

Benefit: Can think about models that are not the simple linear models used in regression settings.
Cost: need to make more assumptions about the distribution of the error term.
Given that choice, can estimate a much wider range of estimation problems.

Intuition about construction

Setup

y : data

θ : parameters

Likelihood function:
$L(y, \theta)$: "How likely we are to have observed y as a function of the parameters."
In the applications we are going to look at, the observations will be independent, and we can write the likelihood function as

$$L(y, \theta) = \prod_{t=1}^{T} L_t(y_t, \theta)$$

where
y_t is observation number t.
$L_t(y_t, \theta)$ is the probability distribution of y_t.

As a rule we can work with the log of the likelihood function, instead of the likelihood function directly

- A max of one will be a max of the other
- The log is typically much easier to find a max of.

Let

$$\ell(y) = \log L(y, \theta)$$

Since

$$L(y, \theta) = \prod_{t=1}^{T} L_t(y_t, \theta)$$

$$\ell(y) = \log L(y, \theta) = \log \left(\prod_{t=1}^{T} L_t(y_t, \theta) \right) = \sum_{t=1}^{T} \log L_t(y_t, \theta) = \sum_{t=1}^{T} \ell_t($$

Definition: The maximum likelihood estimate is the set of parameters θ that maximizes the value of the likelihood function, or alternatively the log likelihood function.

$$\hat{\theta}^{ml} = \arg\max_{\theta} \ell(y, \theta)$$

or

$$\ell(y, \hat{\theta}^{ml}) \geq \ell(y, \theta) \; \forall \; \theta \in \Theta$$

An alternative formulation can be found by looking at the first order conditions for a maximum of the likelihood function.

$$\frac{\partial}{\partial \theta}\ell(y,\theta) = \frac{\partial}{\partial \theta}\sum_{t=1}^{T}\ell_t(y_t,\theta) = \sum_{t=1}^{T}\frac{\partial}{\partial \theta}\ell_t(y_t,\theta) = 0$$

These give two definitions of how to find a ML estimate

- The max of the loglikelihood function: Type I.
- The First Order Condition for a max of the log likelihood function: Type II.

General about Maximum Likelihood

It can be shown that under the assumed probability assumption being correct, maximum likelihood estimators have a number of desirable properties.

1. Any ML estimator is consistent (In large samples it converges to the true parameter.)
2. ML estimators are asymptotically normal (as the number of observations increase, they move towards the normal distribution.)
3. ML estimators are asymptotically efficient. (As the number of observations increase, the ML estimators achieve the so called Cramér-Rao lower bound, which is the minimum possible covariance matrix for an unbiased estimator.
4. Once the probability distribution is specified and the problem is set up, ML estimators are straightforward to implement as nonlinear optimization problems, and will be easy to solve on a computer.

The ML estimators thus have a number of desirable properties, as well as being easy to work with. For example, the usual test statistics, based on the Wald, LM and LR principles, are easily accessible.

Let us look at the LR statistic:

Letting θ be the parameters, and $\mathbf{X}$ the data, $L(\theta, \mathbf{X})$ is the likelihood function. We want to compare the fit of an unrestricted estimate, let us call that $\hat{\theta}$, to a restricted estimate $\tilde{\theta}$. The restricted estimate $\tilde{\theta}$ is found by minimizing the likelihood function imposing the restrictions.

The LR statistic is calculated as

$$LR = 2 \ln \left(\frac{L(\hat{\theta}, \mathbf{X})}{L(\tilde{\theta}, \mathbf{X})} \right)$$

(This is where the name likelihood ratio is coming from, it is the ratio of two likelihoods.

Computational device

Even if one has problems with the swallowing the assumed distributional assumption, the ML method is still a useful *computational* device, it allows calculation of estimates in situations where it would be very hard to get an estimator any other way.

ML estimation of binomial variable

We are observing outcomes y_t from a binomial distribution

$$y_t = \begin{cases} a & \text{with probability} \quad p \\ b & \text{with probability} \quad 1-p \end{cases}$$

1. Determine the Maximum Likelihood estimator of p.

ML estimation of binomial variable - Solution

The inference problem is to estimate the probability p from a sample of T observation of y, $\{y_t\}_{t=1}^{T}$.
Suppose we observe n outcomes of $y_t = a$, and $(T - n)$ outcomes of $y_t = b$.
The "probability" of observing this outcome for a given p is

$$p^n(1-p)^{T-n}$$

To find the maximum likelihood estimator we will maximize this with respect to p, the parameter of interest.
Formally, ML proceeds by creating a likelihood function L, a function of the data (y) and parameters (p).

In this case the *likelihood* function is

$$L(y, p) = p^n (1 - p)^{T-n}$$

This likelihood function is to be maximized with respect to p, the parameter.
In practice we often work with an equivalent formulation, and take logs to get the *log-likelihood* function

$$\begin{aligned} \ell(y, p) &= \log L(y, p) \\ &= n \log(p) + (T - n) \log(T - n) \end{aligned}$$

A maximum for this log-likelihood function is also a maximum for the likelihood function, but it is more easy to work with.

The first order condition for a maximum of the log-likelihood function is

$$\frac{\partial}{\partial p}\ell(y,p) = n\frac{1}{p} - (T-n)\frac{1}{1-p}$$

set this equal to zero and solve for p

$$n\frac{1}{p} - (T-n)\frac{1}{1-p} = 0$$

$$n(1-p) = (T-n)p$$

$$n - np = Tp - np$$

$$n = Tp$$

$$p = \frac{n}{T}$$

Thus, the Maximum Likelihood estimator of p, $\widehat{p}^{ml}$, is

$$\widehat{p}^{ml} = \frac{n}{T}$$

ML estimation of binomial variable - using R

y_t follows a binomial distribution

$$y_t = \begin{cases} a & \text{with probability} \quad p \\ b & \text{with probability} \quad 1-p \end{cases}$$

1. Set $p = 0.5$, simulate a number of outcomes, and estimate the model using ML.

ML estimation of binomial variable - Solution

Suppose we observe n outcomes of $y_t = a$, and $(T - n)$ outcomes of $y_t = b$.
The "probability" of observing this outcome for a given p is

$$p^n(1-p)^{T-n}$$

To find the maximum likelihood estimator we will maximize this with respect to p, the parameter of interest.
Formally, ML proceeds by creating a likelihood function L, a function of the data (y) and parameters (p).

In this case the *likelihood* function is

$$L(y, p) = p^n (1 - p)^{T-n}$$

This likelihood function is to be maximized with respect to p, the parameter.
In practice we often work with an equivalent formulation, and take logs to get the *log-likelihood* function

$$\begin{aligned} \ell(y, p) &= \log L(y, p) \\ &= n \log(p) + (T - n) \log(T - n) \end{aligned}$$

```
loglik <- function (p) {
   T <- length(y)
   n <- sum(y)
   ll <- n*log (p) + (T-n)* log(1-p)
   return(ll)
}
y <- c(1,0,1,0,1,0,1,0,1,0,1,0)
library(maxLik)
ml <- maxLik(loglik, start=c(0.25))
summary(ml)
```

Result in

```
> summary(ml)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 4 iterations
Return code 1: gradient close to zero
Log-Likelihood: -8.317766
1  free parameters
Estimates:
     Estimate Std. error t value  Pr(> t)
[1,]  0.50000    0.14434  3.4641 0.000532 ***
```

ML estimation of uniform distrubution

ML estimation of uniform distribution.
A variable y_t is drawn from an uniform distribution on the interval $[0, b]$ if the probability distribution of y_t is

$$p(y_t) = \begin{cases} \frac{1}{b} & \text{if } y_t \in [0, b] \\ 0 & \text{otherwise} \end{cases}$$

1. Determine the maximum likelihood estimator of b.

ML estimation of uniform distribution.
The only unknown parameter to estimate is the value b. Given a sample y_t, by the definition of the distribution we know that

$$b \geq \max_t y_t$$

The likelihood of observing a set of y_t is

$$L(y, b) = \left(\frac{1}{b}\right)^T$$

Note that this problem can not be solved the usual way, since if we take logs and try to solve the first order conditions:

$$\log L = T(\log(1) - \log(b)) = -T \log(b)$$

$$\frac{\partial}{\partial b} = -T\frac{1}{b} = 0$$

or

$$\frac{1}{b} = 0$$

which can not be set equal to zero, but will go towards zero as $b \to \infty$.

Thus, the first order conditions can not be used to find an estimate of b, but from the likelihood function itself

$$L(y, b) = \left(\frac{1}{b}\right)^T$$

it should be obvious that it will have a maximum at the lowest possible b, which in this case is

$$b = \max_t y_t$$

ML estimation of linear regression

Max Likelihood estimation of OLS regression.
Suppose we are given data x_t and outcomes y_t, where the model postulates that y is related to x by

$$y_t = x_t b + u_t,$$

where u_t is some error term.
To do Maximum Likelihood, we need to make distributional assumptions about the error term u_t. The simplest assumption is to make all errors to be independently, independently normally distributed, with mean zero and variance $\sigma^2 < \infty$:

$$u_t \sim N\left(0, \sigma^2\right)$$

1. Determine the Maximum Likelihood estimator of b.
2. Determine the Maximum Likelihood estimator of σ^2.

Max Likelihood estimation of OLS regression.
Recall the distribution function for the normal distribution.

$$f(u_t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2} u_t^2}$$

Replace u_t with $y_t - x_t b$:

$$f(y_t - x_t' b) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2} (y_t - x_t' b)^2}$$

We are interested in estimating the parameters b and σ. Form the *likelihood function* L:

$$L_T(b, \sigma, X_T, Y_T) = \prod_{t=1}^{T} f(y_t - x_t' b)$$

we include the data $X_T = \{x_1, \cdots, x_T\}$ and $Y_T = \{y_1, \cdots, y_T\}$ in the arguments to make explicit the fact that the likelihood function is also a function of the observed data.

We find the ML estimates from

$$b_T^{ml} = \arg\max_b L_T(b, \sigma, X_T, Y_T)$$

$$\sigma_T^{ml} = \arg\max_\sigma L_T(b, \sigma, X_T, Y_T)$$

Intuitively, by this maximisation we find the parameters b and σ that make the observations $x_1, \cdots, x_T$ *most likely* to have happened.

Let us calculate the explicit estimates. It is easier to find the maximum of the *log-likelihood function*.

$$
\begin{aligned}
\ell_T &= \ell_T(b, \sigma, X_T, Y_T) \\
&= \ln L_T(b, \sigma, X_T, Y_T) \\
&= \ln \left(\prod_{t=1}^{T} f(y_t - x_t' b) \right) \\
&= \sum_{t=1}^{T} \ln \left(f(y_t - x_t' b) \right) \\
&= -\sum_{t=1}^{T} \ln \left(\frac{1}{\sigma} \right) - \sum_{t=1}^{T} \ln \left(\frac{1}{\sqrt{2\pi}} \right) - \sum_{t=1}^{T} \frac{1}{2} \frac{1}{\sigma^2} \left(y_t - x_t' b \right)^2
\end{aligned}
$$

We use the first order conditions:

$$\frac{\partial \ell_T}{\partial b} = \frac{1}{2}\frac{1}{\sigma}2\sum_{t=1}^{T} x_t(y_t - x_t'b) = 0$$

$$\frac{\partial \ell_T}{\partial \sigma^2} = -\sum_{t=1}^{T}\frac{1}{\sigma} - \sum_{t=1}^{T}\frac{1}{2}\left(y_t - x_t'b\right)^2\left(-\frac{2}{\sigma^3}\right) = 0$$

Solve for b:

$$\sum_{t=1}^{T} y_t x_t - \sum_{t=1}^{T} x_t x_t'b = 0$$

$$\left[\sum_{t=1}^{T} x_t y_t\right] = \left[\sum_{t=1}^{T} x_t x_t'\right] b$$

$$\hat{b}_T^{ml} = \left[\sum_{t=1}^{T} x_t x_t'\right]^{-1}\left[\sum_{t=1}^{T} x_t y_t\right]$$

Solve for σ^2:

$$\frac{1}{\sigma}\sum_{t=1}^{T}(-1) + \frac{1}{\sigma^3}\sum_{t=1}^{T}\left(y_t - x_t'b\right)^2 = 0$$

$$-T\sigma^2 + \sum_{t=1}^{T}(y_t - x_t'b)^2 = 0$$

$$\hat{\sigma}^2_{ml} = \frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\hat{b}^2_{ml})^2$$

Note that $\hat{b}^{ml}_T$ in this case is the same as the OLS estimate. This will in general not be the case. The two are derived under different assumptions.

Max Likelihood estimation of OLS regression.
Consider the model

$$y_t = a + bx_t + u_t,$$

where u_t is some error term. Suppose the constant $a = 2$ and $b = 2$, and the error term is normally distributed with mean 0 and variance 1. Simulate 100 observations of this model, and show the estimation of the model using Maximum Likelihood.

Max Likelihood estimation of OLS regression.
Recall the distribution function for the normal distribution.

$$f(u_t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}u_t^2}$$

Replace u_t with $y_t - a + bx_t$:

$$f(y_t - x_t'b) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(y_t - a - bx_t)^2}$$

We are interested in estimating the parameters b and σ. Form the *likelihood function* L:

$$L_T(b, \sigma, X_T, Y_T) = \prod_{t=1}^{T} f(y_t - a - bx_t)$$

As a rule, it is easier to find the maximum of the *log-likelihood function*.

$$
\begin{aligned}
\ell_T &= \ell_T(b, \sigma, X_T, Y_T) \\
&= \ln L_T(b, \sigma, X_T, Y_T) \\
&= \ln \left(\prod_{t=1}^{T} f(y_t - a - bx_t) \right) \\
&= \sum_{t=1}^{T} \ln \left(f(y_t - a - bx_t) \right) \\
&= -\sum_{t=1}^{T} \ln \left(\frac{1}{\sigma} \right) - \sum_{t=1}^{T} \ln \left(\frac{1}{\sqrt{2\pi}} \right) - \sum_{t=1}^{T} \frac{1}{2} \frac{1}{\sigma^2} (y_t - a - bx_t)^2
\end{aligned}
$$

We apply this log likelihood function directly to the R maximum likelihood routine.
First, the simulation of the model. The form of the X variable was not specified, so let us use the integers from 1 to 100.

```
a <- 2
b <- 2
sigma <- 1
N <- 100
x <- 1:N
sigma <-1
y <- a + b*x + rnorm(N,0,sigma)
```

Then, ml estimation. We first need to write the likelihood function as a R function.

```
loglik <- function(param) {
   N=length(x)
   alpha <- param[1]
   beta  <- param[2]
   sigma <- param[3]
   e <- y - ( alpha + beta*x )
   ll <- -0.5 * N * log(2*pi) - N*log(sigma) - sum(0.5*(e)
   return(ll)
}
```

This is then feed to the ML implementation in the library `maxLik`

```
library(maxLik)
ml <- maxLik(loglik, start=c(1,1,1))
summary(ml)
```

With output

```
> summary(ml)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 15 iterations
Return code 1: gradient close to zero
Log-Likelihood: -141.5555
3  free parameters
Estimates:
      Estimate Std. error  t value   Pr(> t)
[1,] 1.9069817  0.2009801   9.4884 < 2.2e-16 ***
[2,] 2.0013569  0.0034545 579.3429 < 2.2e-16 ***
[3,] 0.9966221  0.0704751  14.1415 < 2.2e-16 ***
--------------------------------------------
```

Summarizing Maximum Likelihood estimation

Starting point: The underlying probability distribution that generated the data.
Powerful: the whole distribution has potentially more information than "minimizing distance"
Potential problem: ML is always dependent on the specified probability distribution being "close to correct"
Some important examples of estimation problems where estimation is done using maximum likelihood.

- Limited dependent variable models (Probit/Logit)
- ARCH
- VARs
- Factor analysis