Maximum Likelihhod

Bernt Arne Ødegaard

1 The method of Maximum Likelihood.

An alternative paradigm that is used to generate estimation methods.

Consider how one develop the least squares estimator. In that development no mention is made of probabilities. All that was used was to minimize the distance between the predicted linear regression and the observed data. It was only when one either assumed normality or appealed to large sample results we could come up with results about distributions of the OLS estimator.

In Maximum Likelihood we start in the opposite end. We start by making probability assumptions, we assume we know exactly the probability distribution that made the observed data "most likely" to have been observed.

As soon as we formulate the problem this way, it is clear that we can think about models that are not the simple linear models which we work with in regression settings.

The cost of using Maximum Likelihood is that we need to make more assumptions about the distribution of the error term, but if we are willing to make these, we can estimate a much wider range of estimation problems.

1.1 Intuition about construction

Setup

$$y: \text{ data}$$

$$\theta: \text{ parameters}$$

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

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

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

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.

1.2 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.

1.3 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.

2 Cases

2.1 Maximum likelihood estimation of binomial distribution

Exercise 1.

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.

Solution to Exercise 1.

ML estimation of binomial variable.

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

$$\begin{aligned} n\frac{1}{p} - (T-n)\frac{1}{1-p} &= 0 \\ n(1-p) &= (T-n)p \\ n - np &= Tp - np \\ n &= Tp \\ p &= \frac{n}{T} \end{aligned}$$

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

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

Exercise 2.

ML estimation of binomial variable.

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.

Solution to Exercise 2.

ML estimation of binomial variable.

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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

2.2 ML estimation of exponential

Exercise 3.

ML estimation of exponenentially distributed variables.

Suppose you observe y_t which has an exponential distribution. Further assume they are independent, which means that the probability density function for each y_t is

$$f(y_t) = \theta e^{-\theta y_t}$$

1. Given a sample of observations, $\mathbf{y} = \{y_t\}, t = 1, 2, \cdots, T$, determine the Maximum Likelihood estimate of the parameter θ.

Solution to Exercise 3.

ML estimation of exponenentially distributed variables.

The likelihood function is

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

Take logs,

$$\ell(\theta, \mathbf{y}) = \log L(\theta, \mathbf{y}) = \sum_{t=1}^{T} \log(\theta) - \theta y_t$$

First order condition

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

Solve for θ:

$$\sum_{t=1}^{T} \frac{1}{\theta} = \sum_{t=1}^{T} y_t$$

$$T \frac{1}{\theta} = \sum_{t=1}^{T} y_t$$

$$\frac{1}{\theta} = \frac{\sum_{t=1}^{T} y_t}{T}$$

$$\theta = \frac{T}{\sum_{t=1}^{T} y_t}$$

and

$$\widehat{\theta}^{ml} = \frac{T}{\sum_{t=1}^{T} y_t}$$

3 ML estimation of Poisson distribution

Exercise 4.

ML Poisson [5]

A variable y has a Poisson distribution if

$$p(y=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

λ is a parameter that is to be estimated.

1. Find the Maximum Likelihood estimator for λ.

Solution to Exercise 4.

ML Poisson [5]

The likelihood function for a series of observations $y_t, t = 1, 2, \cdots, T$ is

$$L = \prod_{t=1}^{T} \frac{\lambda^{y_t}}{y_t!} e^{-\lambda}$$

Take logs:

$$\begin{aligned} ł &= \log L \\ &= \sum_{t=1}^{T} \{\log(\lambda^{y_t}) - \log(y_t!) - \lambda\} \\ &= \sum_{t=1}^{T} \{y_t \log(\lambda) - \log(y_t!) - \lambda\} \end{aligned}$$

Take first derivatives to solve the first order conditions for a stationary point

$$\frac{\partial}{\partial \lambda} = \sum_{t=1}^{T} \left\{ y_t \frac{1}{\lambda} - 1 \right\} = 0$$

$$\rightarrow \frac{1}{\lambda} \left(\sum_{t=1}^{T} y_t \right) = T$$

$$\rightarrow \frac{1}{T} \sum_{t=1}^{T} y_t = \lambda$$

The maximum likelihood estimator is

$$\widehat{\lambda}_T^{ml} = \frac{\sum_{t=1}^{T} y_t}{T}$$

3.1 ML estimation of uniform distrubution

Exercise 5.

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.

Solution to Exercise 5.

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

$$b = \max_t y_t$$

4 ML estimation of normal distribution

Exercise 6.

Maximum likelihood estimation of normally distributed variables.
Suppose a variable x_i has a normal distribution with mean μ and variance σ^2.

1. Determine the maximum likelihood estimator of μ.
2. Determine the maximum likelihood estimator of σ^2.

Solution to Exercise 6.

Maximum likelihood estimation of normally distributed variables.
First recall the probability distribution for a normally distributed varible x_i

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

The likelihood function is

$$L(x;\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

We will instead of the likelihood function maximize the log-likelihood function:

$$\ell(y;\theta) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} -\frac{1}{2}\frac{1}{\sigma^2}(x_i-\mu)^2$$

Rewrite the log-likelihood function as

$$\ell(y;\theta) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \sum_{i=1}^{n} -\frac{1}{2}\frac{1}{\sigma^2}(x_i-\mu)^2$$

We find the estimator from the first order conditions, first estimating μ

$$\frac{\partial \ell}{\partial \mu} = \sum_i 2\frac{1}{2\sigma^2}(y_i-\mu) = 0$$

$$\rightarrow \sum_i (y_i-\mu) = 0$$

$$\rightarrow \sum_i y_i = n\mu$$

$$\widehat{\mu}^{ml} = \frac{\sum_i y_i}{n}$$

and then estimating σ^2.

$$\frac{\partial \ell}{\partial \sigma} = \sum_{i=1}^{n} -\frac{1}{\sigma} - \sum_{i=1}^{n} \frac{1}{2}\left(\frac{0-(x_i-\mu)^2 2\sigma}{\sigma^4}\right)$$

$$0 = \sum_{i=1}^{n} -\frac{1}{\sigma} + \sum_{i=1}^{n} \frac{1}{2}\left(\frac{2(x_i-\mu)^2\sigma}{\sigma^4}\right)$$

$$0 = \sum_{i=1}^{n} -\frac{1}{\sigma} + \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{\sigma^3}\right)$$

$$0 = -\sum_{i=1}^{n} 1 + \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$0 = -n + \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{\sigma^2}\right)$$

$$n = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\hat{\sigma}^2_{ml} = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

Exercise 7.

Maximum likelihood estimation of normally distributed variables.

Suppose a variable x_i has a normal distribution with mean μ and variance σ^2.

1. Simulate 1000 normally distributed variables with mean 1 and variance 2. Estimate the model using maximum likelihood.

Solution to Exercise 7.

Maximum likelihood estimation of normally distributed variables.

First recall the probability distribution for a normally distributed varible x_i

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

The likelihood function is

$$L(x;\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

We will instead of the likelihood function maximize the log-likelihood function:

$$\ell(y;\theta) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} -\frac{1}{2}\frac{1}{\sigma^2}(x_i-\mu)^2$$

Rewrite the log-likelihood function as

$$\ell(y;\theta) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \sum_{i=1}^{n} -\frac{1}{2}\frac{1}{\sigma^2}(x_i-\mu)^2$$

```
library(maxLik)
# estimate mean and variance of random normal
loglik <- function(param) {
   mu <- param[1]
   sigma <- param[2]
   N <- length(x)
   ll <- -0.5 * N * log(2*pi) - N*log(sigma) - sum(0.5*(x-mu)^2/sigma^2)
   return(ll)
}
x <- rnorm(1000,1,2)
ml <- maxLik(loglik, start=c(0,1))
summary(ml)
```

Results in

```
> summary(ml)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 8 iterations
Return code 1: gradient close to zero
Log-Likelihood: -2086.565
2  free parameters
Estimates:
     Estimate Std. error t value   Pr(> t)
[1,] 1.067690   0.061628  17.325 < 2.2e-16 ***
[2,] 1.949605   0.043615  44.700 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4.1 ML estimation of linear regression

Exercise 8.

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.

Solution to Exercise 8.

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.

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

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

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:

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

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

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

Note that $\hat{b}_T^{ml}$ 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.

Exercise 9.

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.

Solution to Exercise 9.

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*.

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

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)^2/sigma^2)
   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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

5 Background

6 The thory of Maximum Likelihood Estimation (MLE)

I will give an overview of the basic underlying theory of Maximum Likelihood. I will be relying on the book by Cramer (1986), which I find to be a good introduction to the theory. I will use the same notation as that book.

6.1 Definitions

The basic setup consists of observations of the exogenous variable $\mathbf{x}$ and the endogenous variable $\mathbf{y}$, a vector of l parameters θ, and a probability density function

$$p_y(\omega, \theta_0, x) \quad \theta \in \Theta$$

Roughly, this is the probability of observing the outcome $\mathbf{y}$ given observed $\mathbf{x}$ and $\theta = \theta_0$, where θ_0 is the true, but unknown values of the parameters.

A key ingredient of any ML problem is the likelihood function

$$L = L(\theta, \mathbf{y}, \mathbf{x}) = p_y(y, \theta, \mathbf{x})$$

In most cases we will find it easier to deal with the *log-likelihood function*:

$$\ell = \ell(\theta, \mathbf{y}, \mathbf{x}) = \log L(\theta, \mathbf{y}, \mathbf{x}) = \log p_y(y, \theta, \mathbf{x})$$

In the development here we assume L is differentiable.

We also make an assumption for convenience in the development here, namely that $p(\omega, \theta, x)$ is *regular*. I will not give the mathematical definition of regularity, but a key implication is that you can "differentiate through" the integral, that is

$$E\left[\frac{\partial}{\partial\theta} p(\omega, \theta, x)\right] = \frac{\partial}{\partial\theta} E\left[p(\omega, \theta, x)\right]$$

and

$$E\left[\frac{\partial^2}{\partial\theta^2} p(\omega, \theta, x)\right] = \frac{\partial^2}{\partial\theta^2} E\left[p(\omega, \theta, x)\right]$$

For any regular probability distribution, we can find a couple of useful properties:

$$\int p(\omega, \theta_0, x) d\omega = 1 \quad \text{(since } p(\cdot) \text{ is a probability).}$$

$$\rightarrow \int p'(\omega, \theta_0, x) d\omega = 0$$

$$\rightarrow \int p''(\omega, \theta_0, x) d\omega = 0$$

Since we take expectations by integrating over the true parameters:

$$\begin{aligned} E\left[\frac{p'(y, \theta, x)}{p(y, \theta, x)}\right]_{\theta=\theta_0} &= \int \left[\frac{p'(\omega, \theta, x)}{p(\omega, \theta, x)}\right]_{\theta=\theta_0} p(\omega, \theta_0, x) d\omega \\ &= \int \frac{p'(\omega, \theta_0, x)}{p(\omega, \theta_0, x)} p(\omega, \theta_0, x) d\omega \\ &= \int p'(\omega, \theta_0, x) \\ &= 0 \quad \text{by the above result.} \end{aligned}$$

and

$$\begin{aligned} E\left[\frac{p''(y,\theta,x)}{p(y,\theta,x)}\right]_{\theta=\theta_0} &= \int \left[\frac{p''(\omega,\theta,x)}{p(\omega,\theta,x)}\right]_{\theta=\theta_0} p(\omega,\theta_0,x)d\omega \\ &= \int \frac{p''(\omega,\theta_0,x)}{p(\omega,\theta_0,x)} p(\omega,\theta_0,x)d\omega \\ &= \int p''(\omega,\theta_0,x) \\ &= 0 \quad \text{by the above result.} \end{aligned}$$

Define the *score vector*:

$$\begin{aligned} q(\theta) &= \left[\frac{\partial}{\partial\theta}\log L\right] \\ &= \frac{\partial}{\partial\theta} p(y,\theta,x) \\ &= \frac{p'(y,\theta,x)}{p(y,\theta,x)} \end{aligned}$$

Note that under the true $\theta = \theta_0$,

$$E\left[q(\theta_0)\right] = E\left[\frac{p'(y,\theta_0,x)}{p(y,\theta_0,x)}\right] = 0$$

Define the Hessian matrix of the log likelihood function:

$$\begin{aligned} Q &= \frac{\partial^2}{\partial\theta^2}\log L \\ &= \frac{\partial}{\partial\theta}\left[\frac{p'(y,\theta,x)}{p(y,\theta,x)}\right] \\ &= \frac{p''(y,\theta,x) - p'(y,\theta,x)p'(y,\theta,x)}{\left(p'(y,\theta,x)\right)^2} \\ &= \frac{p''(y,\theta,x)}{p(y,\theta,x)} - \frac{(p'(y,\theta,x))^2}{(p(y,\theta,x))^2} \\ &= \frac{p''(y,\theta,x)}{p(y,\theta,x)} - \frac{p'(y,\theta,x)}{p(y,\theta,x)}\frac{p'(y,\theta,x)}{p(y,\theta,x)} \end{aligned}$$

$$\begin{aligned} E[Q] &= E\left[\frac{p''(y,\theta,x)}{p(y,\theta,x)} - \frac{p'(y,\theta,x)}{p(y,\theta,x)}\frac{p'(y,\theta,x)}{p(y,\theta,x)}\right] \\ &= 0 - E\left[q(\theta)q'(\theta)\right] \\ &= -E\left[q(\theta)q'(\theta)\right] \\ &= -V(\theta), \quad \text{where } V \text{ is the covariance matrix of } \theta. \end{aligned}$$

Define the *Fisher information matrix*

$$H = -Q(\theta_0)$$

$$\rightarrow\ H = V(\theta_0)$$

$$\rightarrow\ H \text{ is positive definite.}$$

6.2 Maximum of likelihood function.

We will now show that the expected log likelihood has a maximum at the true parameters. Define the function

$$\Psi(\theta) = E\left[\log L(\theta)\right]$$

By the regularity of $p(\cdot)$, and hence of $\log L$,

$$\Psi'(\theta) = \frac{\partial}{\partial\theta} E\left[\log L(\theta)\right] = E[q(\theta)]$$

At the true parameters θ_0, we have shown that $E[q(\theta_0)] = 0$, hence

$$\Psi(\theta_0) = 0$$

The true parameters is thus a stationary point for the log likelihood function. To check that it is a maximum, we need to look at the second derivatives.

$$\begin{aligned}
\Psi''(\theta) &= \frac{\partial^2}{\partial\theta^2} E\left[\log L(\theta)\right] \\
&= E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta)\right] \\
&= E[Q(\theta)]
\end{aligned}$$

Hence

$$\Psi''(\theta_0) = -H$$

Since H is positive definite, $-H$ is negative definite, which shows that θ_0 is a local maximum.

Thus, *at the true parameters, the log likelihood function has a maximum.* Let us for now just assume that this is a global maximum.

6.3 The Cramer-Rao inequality

We next show an efficiency result, the famous Cramer-Rao inequality, which we will use to show that any maximum likelihood estimator will asymptotically achieve the lowest possible variance among the set of unbiased estimators.

Let $t = t(y, x)$ be any estimator. We will look at unbiased estimators

$$E[t] = \theta_0$$

or

$$\int t(\omega, x)p(\omega, \theta_0, x)d\omega = \theta_0$$

Derive this with respect to θ on both sides of this expression

$$\begin{aligned}
&\rightarrow \frac{\partial}{\partial\theta}\int t(\omega, x)p(\omega, \theta_0, x)d\omega = \mathbf{I} \\
&\rightarrow \int \frac{\partial}{\partial\theta}\left\{t(\omega, x)p(\omega, \theta_0, x)\right\}d\omega = \mathbf{I} \\
&\rightarrow \int t(\omega, x)\left\{\frac{\partial}{\partial\theta}p(\omega, \theta_0, x)\right\}d\omega = \mathbf{I} \\
&\rightarrow \int t(\omega, x)\left\{p'(\omega, \theta_0, x)\right\}d\omega = \mathbf{I}
\end{aligned}$$

$$\rightarrow \int t(\omega, x) \left\{ \frac{p'(\omega, \theta_0, x)}{p(\omega, \theta_0, x)} p(\omega, \theta_0, x) \right\} d\omega = \mathbf{I}$$
$$\rightarrow E[tq(\theta_0)] = \mathbf{I}$$

In showing the Cramer-Rao result, we use a similar trick to the Gauss-Markov theorem when we show the efficiency result. Consider

$$z = t - \theta_0 - H^{-1} q(\theta_0)$$

Note that

$$E[z] = E[t] - \theta_0 - H^{-1} E[q(\theta_0)] = \theta)0 - \theta_0 - H^{-1} \cdot 0 = 0$$

Calculate the variance of z

$$\begin{aligned}
\text{var}(z) &= E[zz'] \\
&= E[((t-\theta_0) - H^{-1}q(\theta_0))((t-\theta_0) - H^{-1}q(\theta_0))'] \\
&= E[((t-\theta_0)(t-\theta_0)' - H^{-1}q(\theta_0))(t-\theta_0)' \\
&\quad -(t-\theta_0)q(\theta_0))'H^{-1} + H^{-1}q(\theta_0)q(\theta_0)H^{-1}] \\
&= E[((t-\theta_0)(t-\theta_0)')] - E[H^{-1}q(\theta_0)(t-\theta_0)'] \\
&\quad -E[(t-\theta_0)q(\theta_0)'H^{-1}] + H^{-1}E[q(\theta_0)q(\theta_0)']H^{-1}] \\
&= E[((t-\theta_0)(t-\theta_0)')] - H^{-1} \\
&= V(t) - H^{-1}
\end{aligned}$$

Hence

$$V(z) = V(t) - H^{-1}$$

or

$$V(t) = H^{-1} + V(z)$$

Since $V(z)$ is a covariance matrix, H^{-1} is thus a lower bound on the covariance matrix of t, which we assumed to be *any* unbiased estimator. This is the famous inequality of Cramer-Rao, that H^{-1}, where $H = E[q(\theta_0)q(\theta_0)']$, is a lower bound on the variance of any unbiased estimator. Note that this is a stronger result than the Gauss-Markov theorem, since that result limited itself to *linear* unbiased estimators. Here the result covers *any* unbiased estimator.

Hence, if we can show that the covariance matrix of any ML estimator *is* H^{-1}, we have shown that any ML estimator is efficient. Unfortunately, we can not show this result in general, but we *can* show that it will hold asymptotically as the number of observations increase.

6.4 Defining the ML estimator

The ML estimator is defined as

$$\theta_T^{ml} = \arg\max_{\theta} L(\theta, y, x).$$

We can alternatively defined the ML estimator from the first order conditions for a maximum:

$$\left[\frac{\partial}{\partial \theta} \log L \right]_{\theta = \widehat{\theta}} = q(\widehat{\theta}) = 0$$

If the solution is interior,

$$\left[\frac{\partial^2}{\partial \theta^2} \log L \right]_{\theta = \widehat{\theta}} = Q(\widehat{\theta})$$

is negative definite.

6.5 Properties of ML estimators

Consistency

We can show an asymptotic result

$$\widehat{\theta}^{ml} \xrightarrow{P} \theta_0$$

as the number of observations increase.

Let me sketch part of the argument in the simple case where we have *iid* observations. Then

$$L_T(\theta, y, x) = \prod_{t=1}^{T} p(y_t, \theta, x_t)$$

and

$$\log L_T(\theta, y, x) = \sum_{t=1}^{T} \log p(y_t, \theta, x_t)$$

The asymptotics of the case is beyond this course, but we can show that

$$\frac{1}{T} \log L_T(\theta) \xrightarrow{P} E[\log L_T(\theta)]$$

for all θ. But this does not show the consistency of the ML estimator, since this is

$$\theta_T^{ml} = \arg\max_{\theta} \log L_T(\theta, y, x).$$

We need to use a stronger law of large numbers to show that

$$\arg\max_{\theta} \log L_T(\theta, y, x) \xrightarrow{P} \arg\max_{\theta} \log L_T(\theta_0, y, x)$$

and further show that this implies that

$$\theta_T^{ml} \xrightarrow{P} \theta_0$$

Let me just assert that such a LLN can be applied, implying the consistency of the ML estimator.

Asymptotic normality.

I will just sketch some of the steps in the proof of asymptotic normality.

Consider the score function $q(\theta)$, the vector of first derivatives. Write this as a (first order) Taylor expansion around the true parameter vector θ_0:

$$q_n(\theta) \approx q_n(\theta_0) + \left[\frac{\partial}{\partial \theta} \widehat{q}_n(\theta_0)\right] (\widehat{\theta}_n - \theta_0)$$

$$q_n(\theta) \approx q_n(\theta_0) + \widehat{Q}_n(\theta_0)(\widehat{\theta}_n - \theta_0)$$

$$\rightarrow (\widehat{\theta}_n - \theta_0) = \left[\widehat{Q}_n(\theta_0)\right]^{-1} \left(q_n(\widehat{\theta}) - q_n(\theta_0)\right)$$

Since $q_n(\widehat{\theta}) = 0$

$$(\widehat{\theta}_n - \theta_0) = \left[-\widehat{Q}_n(\theta_0)\right]^{-1} q_n(\theta_0)$$

$$\rightarrow (\widehat{\theta}_n - \theta_0) = \left[-\frac{1}{n}\widehat{Q}_n(\theta_0)\right]^{-1} \frac{1}{n} q_n(\theta_0)$$

$$\rightarrow \sqrt{n}(\widehat{\theta}_n - \theta_0) = \left[-\frac{1}{n}\widehat{Q}_n(\theta_0)\right]^{-1} \frac{1}{\sqrt{n}} q_n(\theta_0)$$

To show the convergence properties of this, would split into two parts:

1.
$$\left[-\frac{1}{n}\widehat{Q}_n(\theta_0)\right] \xrightarrow{P} E[-Q(\theta_0)] = H$$
$$\rightarrow \left[-\frac{1}{n}\widehat{Q}_n(\theta_0)\right]^{-1} \xrightarrow{P} H^{-1}$$

2.
$$\frac{1}{\sqrt{n}}q_n(\theta_0)$$
converges in distribution to a normal random variable with asymptotic covariance matrix
$$E[q(\theta_0)q(\theta_)] = H$$

Thus, the covariance matrix of
$$\left[-\frac{1}{n}\widehat{Q}_n(\theta_0)\right]^{-1}\frac{1}{\sqrt{n}}q_n(\theta_0)$$
will converge to
$$H^{-1}HH^{-1} = H^{-1}$$

This incidentally also shows the asymptotic efficiency of the ML estimator, since its covariance matrix converges to the Cramer-Rao lower bound H^{-1}.

6.6 Summarizing

We have thus shown what arguments will be used to show 3 important properties of any maximum likelihood estimator.

- Consistency
- Asymptotic normality
- Asymptotic efficiency

A fourth property that is important is

- Computability.

The computation of an ML estimator is a matter of maximizing the likelihood, which is a well understood optimization problem.

Also, the 3 classical test statistics are easy to compute in the ML context.

1. Likelihood ratio statistic.
2. Lagrange multiplier statistic (Rao's score test.)
3. Wald test.

7 Summarizing Maximum Likelihood

Let us now summarize Maximum Likelihood estimation.

Its starting point is the underlying probability distribution that generated the data.

This is the source of the power of maximum likelihood, but also of the problem of maximum likelihood, namely that it is always dependent on the specified probability distribution.

With that qualification we have in ML a very powerful tool that allows us to identify estimation problems we could not have attempted in regression settings.

Some important examples of estimation problems where estimation is done using maximum likelihood.

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

8 Readings

The basics of the theory of Maximum Likelihood is covered in Cramer (1986).

Alternatively, look at (Davidson and MacKinnon, 1993, Chapter 8)

9 Further Readings

Maximum Likelihood compared to other estimation methods: (Rao, 1973, 5d)

References

J S Cramer. *Econometric applications of Maximum Likelihood methods.* Cambridge University Press, 1986.

Russel Davidson and James G MacKinnon. *Estimation and Interference in Econometrics.* Oxford University Press, 1993.

C Radhakrisna Rao. *Linear Statistical Inference and its applications.* Wiley, Second edition, 1973.