Elements of Likelihood Theory

We will now examine elements of the likelihood theory, regarding properties of the resulting maximum likelihood estimator (MLE), and testing procedures based on the likelihood functions like the likelihood ratio test (LR), and the Score test.

The likelihood function is definable in parametric statistical models (even though the Empirical likelihood

function, briefly considered in the notes regarding GMM, is a non-parametric adaptation).

Hence in this framework, the sample z_n, is paired with a parametric model, $\{P_\theta, \theta \in \Theta\}$, where P_θ is a distribution related to the sample at scenario $\theta \in \Theta$.

The mapping $\theta \to P_\theta$ is 1-1, hence the parameter value θ completely specifies the distribution in the model that indexes.

E.g. 1 $Z_n = (Z_{(1)}, Z_{(2)}, \dots, Z_{(n)})$ is iid and $Z_{(1)} \sim \text{Exp}(\lambda_0)$ distribution with λ_0 unknown.

Remember that the $\text{Exp}(\lambda)$ distribution - $\lambda > 0$ - is the one with density

$$f(z) = \begin{cases} 0, & z < 0 \\ \lambda e^{-\lambda z}, & z \geq 0 \end{cases}$$

A natural parametric model emerging from the above information is:

$$\{\text{Exp}(\lambda), \lambda > 0\}$$

There $\Theta = \Lambda$, $\Theta = (0, +\infty)$, $k = 1$, and each value of λ uniquely determines the

relevant exponential distribution. ∎

E.g. 2.

$Z_n = (Z_{(1)}, Z_{(2)}, \dots, Z_{(n)})$ is iid and $Z_{(1)} \sim N(\mu_0, v_0)$ distribution with $\theta_0 := (\mu_0, v_0)$ unknown. This information uniquely determines the following Gaussian model:

$$\left\{ Z_n \sim N\left(\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, v I_{n\times n} \right), \mu \in \mathbb{R}, v > 0 \right\}$$

Here $\theta = (\mu, v)$, $\Theta = \mathbb{R} \times \mathbb{R}_{+}$, $k = 2$ and each value of θ uniquely determines the relevant

Gaussian distribution. ▤

Eg. 3 [Gaussian LS]

$$z_n = [Y_n, X_n],$$

$$Y_n \,|\, \mathcal{G}(X_n) \sim N(X_n \theta_0, I_{n\times n}),$$

with $\theta_0 \in \mathbb{R}^k$ unknown. This information uniquely determines the model

$$\{ Y_n \,|\, \mathcal{G}(X_n) \sim N(X_n \theta, I_{n\times n}), \theta \in \mathbb{R}^k \}$$

where $\Theta = \mathbb{R}^k$, etc. ▤

[In what follows we will investigate further examples]

Note: Remember that we work under the universal assumption of correct specification; $\mathbb{P}_{\theta_0} \in \text{Model}$. We will provide some misspecification considerations later.

Question: Is it possible to construct a criterion C_n that fully and faithfully represent the information that appears in the model jointly with the sample?

Regularity Assumption We will assume that for any $\theta \in \Theta$, $\mathbb{P}_\theta$ has a density function (it should be a real function that depends on (θ, z), say $f(\theta, z)$.

The constructions that follow are easily extendable to cases where densities do not exist.

Kullback-Leibler Divergence from P_{θ_0} and the likelihood function.

Given the Brevity Assumption, for any $\theta \in \Theta$, the KL divergence between the latent P_{θ_0} and P_θ, is defined as:

$$KL(P_{\theta_0}, P_\theta) = \int \ln \left(\frac{f(\theta_0, z)}{f(\theta, z)} \right) f(\theta_0, z)\, dz$$

this lies in the space where Z_n assumes its values with prob. 1

It is not difficult to prove that

$\operatorname{argmin}_{P_\theta \in \text{Model}} KL(P_{\theta_0}, P_\theta) = P_{\theta_0}$, hence the latent distribution has a variational characterization. It is however impossible to recover P_{θ_0}, by minimizing $KL(P_{\theta_0}, P_\theta)$ since this depends on P_{θ_0}. If however the empirical joint distribution of the sample P_n^* (remember P_n^* is the discrete distribution that places probability mass at z_n)

"Approximates" $\mathbb{P}_{\theta_0}$, and the integral that defines $KL(\mathbb{P}_{\theta_0}, \mathbb{P}_\theta)$ is approximated by an empirical mean w.r.t. $\mathbb{P}_n$ we obtain

$$\ln \frac{1}{f(\theta, z_n)} \cdot 1 =$$

$$= \boxed{-\ln f(\theta, z_n)}$$

↳ this can be perceived as an empirical approximation of $KL(\mathbb{P}_{\theta_0}, \mathbb{P}_\theta)$ for any $\theta \in \Theta$.

Hence, we may hope that we can approximate θ_0 (and hence $\mathbb{P}_{\theta_0}$), via

minimizing (w.r.t. θ) the empirical approximation.

Notice that minimizing this is equivalent to maximizing its opposite, whereas maximization is covariant to multiplying this by $1/n$ "for averaging purposes". Hence we arrive at the concept of the (average log-) likelihood function [following the general convention we will denote it with ℓ_n instead of G_n).

Definition Given the Model, the Identifiability Assumption, and if $f(\theta, z_n) > 0$ at the support of $\mathbb{P}_\theta$ (the "smallest" closed subset of the space at which z_n attains its values of probability one) the average log-likelihood function is:

$$\ell_n(\theta) = \frac{1}{n} \ln\left(f(\theta, z_n)\right)$$

E.g. 1 We have that $f(\lambda, z_n) =$

$$= \begin{cases} 0, & \exists i : z_{(i)} < 0 \\ \prod_{i=1}^{n} \lambda \exp(-\lambda z_{(i)}), & z_{(1)} \geq 0 \ \forall i \end{cases}$$

[the support of z_n is $[0, +\infty)^n$] $\forall 0$

Hence $l_n(\lambda, z_n) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\lambda \exp(-\lambda z_{(i)})\right)$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\ln(\lambda) + \ln(\exp(-\lambda z_{(i)}))\right)$$

$$= \ln(\lambda) - \lambda \frac{1}{n}\sum_{i=1}^{n} z_{(i)}.$$

E.g. 2 We have that

$$f(\theta, z_n) = \frac{(2\pi)^{-n/2}}{v^{n/2}\sqrt{\det I_{nxn}}} \exp\left(-\frac{1}{v}(z_n - \begin{pmatrix}\mu\\ \vdots \\ \mu\end{pmatrix})' I_{nxn}^{-1} (z_n - \begin{pmatrix}\mu\\ \vdots \\ \mu\end{pmatrix})\right)$$

$\det I_{nxn} = 1$

And thereby

$$l_n(\theta) \propto -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln v - \frac{1}{2v}(z_n - \begin{pmatrix}\mu\\ \vdots \\ \mu\end{pmatrix})'(z_n - \begin{pmatrix}\mu\\ \vdots \\ \mu\end{pmatrix})$$

$$= -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln v - \frac{1}{2v}\sum (z_{(i)} - \mu)^2$$

E.g. 3. We have that

$$f(\theta, Y_n / \sigma(X_n)) = \frac{(2\pi)^{-n/2}}{\sqrt{\det I_{nn}}} \exp\left(-\frac{1}{2}(Y_n - X_n\theta)'(Y_n - X_n\theta)\right)$$

And thereby

$$l_n(\theta, Y_n/\sigma(X_n)) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\frac{1}{n}(Y_n - X_n\theta)'(Y_n - X_n\theta)$$

$\Downarrow$

the LS criterion.

The MLE

Setting $c_n := -l_n$ and following the general theory the OE Maximum Likelihood Estimator is established:

$\theta_n \in \underset{\theta \in \Theta}{\operatorname{argmin}}\, G_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, h_n(\theta)$

MLE

or more generally (allowing for the existence of optimization error)

$$h_n(\theta_n) \geq \sup_{\theta \in \Theta} h_n(\theta) - \varepsilon_n$$

ε_n: optimization error

Note: the existence arguments for the GMLE fall into the ones of our general theory: e.g. Θ convex, and $h_n(\theta)$ concave, or e.g. Θ compact, and $h_n(\theta)$ continuous, suffice for existence.

E.g. 1

Remember that $\Theta = (0, +\infty)$ → $\Theta = \lambda$ here!

and $l_n(\lambda) = \ln(\lambda) - \lambda \frac{1}{n} \sum_{i=1}^{n} x_i$

l_n is smooth, $\frac{\partial l_n(\lambda)}{\partial \lambda} = \frac{1}{\lambda} - \frac{1}{n} \sum_{i=1}^{n} x_i$

and $\frac{\partial^2 l_n(\lambda)}{\partial \lambda^2} = -\frac{1}{\lambda^2} < 0 \quad \forall \lambda \in \Theta$, hence

the function is strictly concave with Θ convex. The maximizer would be unique and Kuhn-Tucker conditions are not needed to locate it as it is an interior point, as we will see:

f.o.c. $\frac{\partial l_n(\lambda)}{\partial \lambda} = 0 \Leftrightarrow \frac{1}{\lambda} - \frac{1}{n} \sum_{i=1}^{n} x_i \Leftrightarrow$

→ the unique critical point

$$\lambda_n = \frac{n}{\sum_{i=0}^{n} x_{(i)}}$$

→ * Notice that this is almost surely well-defined as $X_{(i)} > 0$ with Prob. 1 $\forall i$.

S.O.C $$\frac{\partial^2 \ell_n(\lambda_n)}{\partial \lambda^2} = -\left(\frac{\sum_{i=0}^{n} x_{(i)}}{n}\right)^2 < 0$$

with Prob. 1, hence the MLE is

$$\hat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} x_{(i)}}$$

→ the reciprocal of the sample Mean.

* It is easily proven that if $X \sim Exp(\lambda_0)$ then $\mathbb{E}(X) = 1/\lambda_0$. We have that

$$\mathbb{E}(\hat{\lambda}_n) = \mathbb{E}\left(\frac{n}{\sum_{i=1}^{n} x_{(i)}}\right) = n\,\mathbb{E}\left(\sum_{i=1}^{n} x_{(i)}\right)^{-1} \neq$$

$$n\left(\mathbb{E}\left(\sum_{i=1}^{n} z_{(i)}\right)\right)^{-1} = n\left(\sum_{i=1}^{n} \mathbb{E}(z_{(i)})\right)^{-1}$$

$X_{(i)} \sim \text{Exp}(\lambda_0), \forall i$

$$= n\left(\sum_{i=1}^{n} 1/\lambda_0\right)^{-1} = n n^{-1} \lambda_0 = \lambda_0$$

Hence the MLE of λ_0 is biased, which implies that the MLE in general is not unbiased (Unbiasedness is a difficult to hold and fragile property!) ∎

E.g. 2 Remember that $\Theta = \mathbb{R} \times \mathbb{R}_{++}$

$$\theta = \begin{pmatrix} \mu \\ v \end{pmatrix}, \quad l_n(\theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln v - \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\frac{(z_{(i)}-\mu)^2}{v}$$

The function is again smooth with

$$\frac{\partial l_n(\theta)}{\partial \theta} = \begin{pmatrix} \partial l_n(\theta)/\partial \mu \\ \partial l_n(\theta)/\partial v \end{pmatrix},$$

$$\frac{\partial \ln(\theta)}{\partial \mu} = \frac{1}{v}\frac{1}{n}\sum_{i=1}^{n}(z_{(i)}-\mu) = \frac{1}{v}\left(\frac{1}{n}\sum_{i=1}^{n} z_{(i)} - \mu\right)$$

$$\frac{\partial \ln(\theta)}{\partial v} = \frac{-1}{2v} + \frac{1}{2v^2}\frac{1}{n}\sum_{i=1}^{n}(z_{(i)}-\mu)^2$$

$$\frac{\partial^2 \ln(\theta)}{\partial\theta\,\partial\theta'} = \begin{pmatrix} \frac{\partial^2 \ln(\theta)}{\partial \mu^2} & \frac{\partial^2 \ln(\theta)}{\partial \mu\,\partial v} \\ \frac{\partial^2 \ln(\theta)}{\partial v\,\partial \mu} & \frac{\partial^2 \ln(\theta)}{\partial v^2} \end{pmatrix}$$

[Second order derivatives are continuous in θ, Young's Th. implies that $\frac{\partial^2 \ln(\theta)}{\partial v\,\partial v_1} = \frac{\partial^2 \ln(\theta)}{\partial v_1\,\partial v}$, check!]

$$\frac{\partial^2 \ln(\theta)}{\partial \mu^2} = \frac{\partial\, 1/2v\,(1/n\sum_{i=1}^{n} z_{(i)} - \mu)}{\partial \mu} = -\frac{1}{v}$$

$$\frac{\partial^2 \ln(\theta)}{\partial v\,\partial v_1} = \frac{\partial\left(1/2v\,(1/n\sum z_{(i)} - \mu)\right)}{\partial v} = -\frac{1}{2v^2}\left(\frac{1}{n}\sum_{i=1}^{n} z_{(i)} - \mu\right)$$

$$\frac{\partial^2 l_n(\theta)}{\partial v^2} = \frac{\partial \left(-\frac{1}{2v} + \frac{1}{2v^2} \frac{1}{n} \sum_{i=1}^{n} (z_{(i)} - \mu)^2 \right)}{\partial v} =$$

$$= \frac{1}{2v^2} - \frac{1}{v^3} \frac{1}{n} \sum_{i=1}^{n} (z_{(i)} - \mu)^2 .$$

We have that:

f.o.c. $\frac{\partial l_n(\theta)}{\partial \theta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Leftrightarrow$

$$\left.\begin{array}{l} \frac{\partial l_n(\theta)}{\partial \mu} = 0 \\ \frac{\partial l_n(\theta)}{\partial v} = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} \frac{1}{v}\left(\frac{1}{n} \sum_{i=1}^{n} z_{(i)} - \mu \right) = 0 \\ -\frac{1}{2v} + \frac{1}{2v^2} \frac{1}{n} \sum_{i=1}^{n} (z_{(i)} - \mu)^2 = 0 \end{array}\right\} \Leftrightarrow$$

$$\left.\begin{array}{l} \mu_n = \frac{1}{n} \sum_{i=1}^{n} z_{(i)} \\ v_n = \frac{1}{n} \sum_{i=1}^{n} (z_{(i)} - \mu_n) \end{array}\right\} \Rightarrow \theta_n = \begin{pmatrix} \mu_n \\ v_n \end{pmatrix}$$

is the critical Point

Notice that:

$$\frac{\partial^2 l_n(\theta_n)}{\partial \mu^2} = -\frac{1}{v_n} < 0 \text{ with Prob 1 since}$$

if $Z_{(i)} \sim N(\mu_0, v_0)$, $\mathbb{P}(Z_{(i)} = 0) = 0$

↓ the distribution has a density

$$\frac{\partial^2 l_n(\theta_n)}{\partial v \partial \mu} = -\frac{1}{v_n}\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} Z_{(i)}}_{\simeq \mu_n} - \mu_n\right) =$$

$$= -\frac{1}{v_n}(\mu_n - \mu_n) = 0$$

$$\frac{\partial^2 l_n(\theta_n)}{\partial v^2} = \frac{1}{2v_n^2} - \frac{1}{v_n^3}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(Z_{(i)} - \mu_n)^2}_{\simeq v_n}$$

$$= \frac{1}{2v_n^2} - \frac{v_n}{v_n^3} = -\frac{1}{v_n^2} < 0 \text{ with Prob 1}$$

Thereby

$$\frac{\partial^2 h(\theta_n)}{\partial \theta \partial \theta'} = \begin{pmatrix} -1/2v_n & 0 \\ 0 & -1/v_n^2 \end{pmatrix}$$

the Hessian of θ_0 is diagonal; hence its eigenvalues are its diagonal elements; hence it is negative definite (remember that we dually defined the MLE via Maximization) iff its diagonal elements are negative; they are with Prob 1.

Hence, with Prob 1, the MLE for θ_0,

$$\theta_n = \begin{pmatrix} \gamma_n \\ v_n \end{pmatrix} = \begin{pmatrix} 1/n \sum_{i=1}^{n} z(i) \\ 1/n \sum_{i=1}^{n} (z(i) - \gamma_n)^2 \end{pmatrix}$$

It is easy to show (do it!) that

$$\mathbb{E}(\gamma_n) = \gamma_0, \quad \mathbb{E}(v_n) = \frac{n-1}{n} v_0, \text{ hence } \mathbb{E}(\theta_n) =$$

$$= \begin{pmatrix} \mathbb{E}(\theta_1) \\ \mathbb{E}(\theta_2) \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \frac{n-1}{n} v_0 \end{pmatrix} \neq \begin{pmatrix} \mu_0 \\ v_0 \end{pmatrix}, \text{ hence}$$

the MLE is biased also in this example:

Notice however that since $\lim_{n\to\infty} \frac{n-1}{n} = 1$,

$$\lim_{n\to\infty} \mathbb{E}(\theta_n) = \begin{pmatrix} \mu_0 \\ v_0 \lim_{n\to\infty} \frac{n-1}{n} \end{pmatrix} = \begin{pmatrix} \mu_0 \\ v_0 \end{pmatrix},$$

hence in this example it is asymptotically unbiased.

Note: In this example no Kuhn-Tucker type conditions to enforce the restriction $v>0$ were needed; the optimizer is an interior point with Prob. 1.

□

E.g. 3

Remember that

$$\ell_n(\theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2n}(Y_n - X_n\theta)'(Y_n - X_n\theta)$$

and due to monotonicity and duality

$$\underset{\theta\in\Theta}{\operatorname{argmax}}\ \ell_n(\theta) \overset{\text{mon}}{=} \underset{\theta\in\Theta}{\operatorname{argmax}}\left(-\frac{1}{2n}(Y_n - X_n(\theta))'(Y_n - X_n\theta)\right)$$

$$\overset{\text{mon}}{=} \underset{\theta\in\Theta}{\operatorname{argmax}}\left(-\frac{1}{n}(Y_n - X_n\theta)'(Y_n - X_n\theta)\right)$$

$$\overset{\text{dual}}{=} \underset{\theta\in\Theta}{\operatorname{argmin}}\ \frac{1}{n}(Y_n - X_n\theta)'(Y_n - X_n\theta)$$

Hence $MLE(\theta_0) = OLSE(\theta_0)$ in this setup. We will try to figure out the reason of this coincidence later. ∎

General Note 1.

Remember that in the construction of the likelihood function $f(\theta, z_n)$ denotes the density, of the (joint) distribution of z_n, at parameter scenario θ. This can be factored as a product of conditional densities (times a marginal) as

$$f(\theta, z_n) = f(\theta, z_{(n)} / z_{(n-1)}, z_{(n-2)}, \dots, z_{(1)}) \times f(\theta, z_{(n-1)} / z_{(n-2)}, z_{(n-3)}, \dots, z_{(1)}) \times \dots \times f(\theta, z_{(1)})$$

where $f(\theta, z_{(n-k)} / z_{(n-k-1)}, z_{(n-k-2)}, \dots, z_{(1)})$ is the density of the conditional distribution of the sample element $z_{(n-k)}$ given $z_{(n-k-1)}, z_{(n-k-2)}, \dots, z_{(1)}$, and $f(\theta, z_{(1)})$ is the marginal density of the first sample element, both at Parameter scenario θ. Hence:

$$l_n(\theta) = \frac{1}{n} \ln\left(f(\theta, z_{(1)})\right) =$$

$$\frac{1}{n} \ln\left(\prod_{i=1}^{n-1} f(\theta, z_{(n-i)} / z_{(n-i)}, \dots, z_{(1)}) \; f(\theta, z_{(1)})\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n-1} \ln f(\theta, z_{(n-i+1)} / z_{(n-i)}, \dots, z_{(1)}) + \frac{1}{n} \ln f(\theta, z_{(1)})$$

Then the part

$$\sum_{i=1}^{n-1} \ln f(\theta, z_{(n-i+1)} | z_{(n-i)}, \ldots, | z_{(1)})$$

is termed as the conditional (part of the) likelihood, while the term

$-\ln f(\theta, z_{(n)})$ is termed as the marginal part of the likelihood. In several models (see e.g. the GARCH(1,1) below) the marginal part is unspecified, while the conditional part is only used for the construction of the estimator.

It is expected that as $n \to \infty$ the marginal part becomes in any case negligible.

- When Z_n is iid then

$$f(\theta, Z_{(n-i+1)} \mid Z_{(n-i)}, \dots, Z_{(1)}) = f(\theta, Z_{(n-i+1)})$$

$\forall i = 1, \dots, n-1$ and thereby

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(\theta, Z_{(i)})$$

As is essentially the case with Eg. 1 and Eg. 2 (what about Eg. 3?).

General Note 2.

The construction of the likelihood function is possible when density functions do not exist; e.g. Z_n is comprised by

discrete random variables, using the at the time probabilistic machinery available.

Example. Z_n is iid, and

$Z_{(1)} \sim Ber(q_0)$ with unknown $q_0 \in (0,1)$.

The natural choice for the statistical model is $\{ Ber^n(q), q \in (0,1) \}$.

Remember that if $Z_{(1)} \sim Ber(q)$ then its probability mass function is

$$p^{(q)}_{Z_{(1)}} : \{0,1\} \to (0,1), \quad p^{(q)}_{Z_{(1)}}(x) = \begin{cases} q, & x=1 \\ 1-q, & x=0 \end{cases}$$

Using an argument based on the definition of the Kullback-Liebler divergence it is provable that in this case the Max functions are usable in place of the non-existent densities. Hence

$$l_n(q) = \frac{1}{n} \sum_{i=1}^{n} \ln p^{(q)}_{z(i)}(z(i)) =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \ln q \cdot 1_{z(i)=1} + \ln(1-q) 1_{z(i)=0}$$

$$= \frac{1}{n} \left[\ln q \cdot \#\{i=1,\dots,n,\ z(i)=1\} + \ln(1-q) \#\{i=1,\dots,n,\ z(i)=0\} \right] = \ln q \cdot (\text{sample freq of } 1) + \ln(1-q)(1 - \text{sample freq of } 1).$$