

# 11 ML Asymptotics. Likelihood Ratio Test.

More rigour: Amemiya (1985)

## 11.0.1 Consistency of ML

Let  $z_i$  be iid with density  $f(z; \theta_0)$  for  $i = 1, \dots, n$ .

$$\hat{\theta}_{ML} = \operatorname{argmax}_{\theta} \sum_{i=1}^n \log f(z_i; \theta)$$

By Khinchine's LLN for any  $\theta$

$$\frac{1}{n} \sum_{i=1}^n \log f(z_i; \theta) \xrightarrow{p} \mathbb{E}_{\theta_0}[\log f(z; \theta)]$$

We can invoke KLLN as given  $z_i$  iid  $\Rightarrow$  any function of  $z_i$  is also iid. We also need to assume the expectation exists. This is taken over the value of the true parameter, but the conditioned  $\theta$  runs across the real line.

### Proposition 11.0.1.

$$\hat{\theta}_{ML} \xrightarrow{p} \operatorname{argmax}_{\theta} \mathbb{E}_{\theta_0}[\log f(z; \theta)]$$

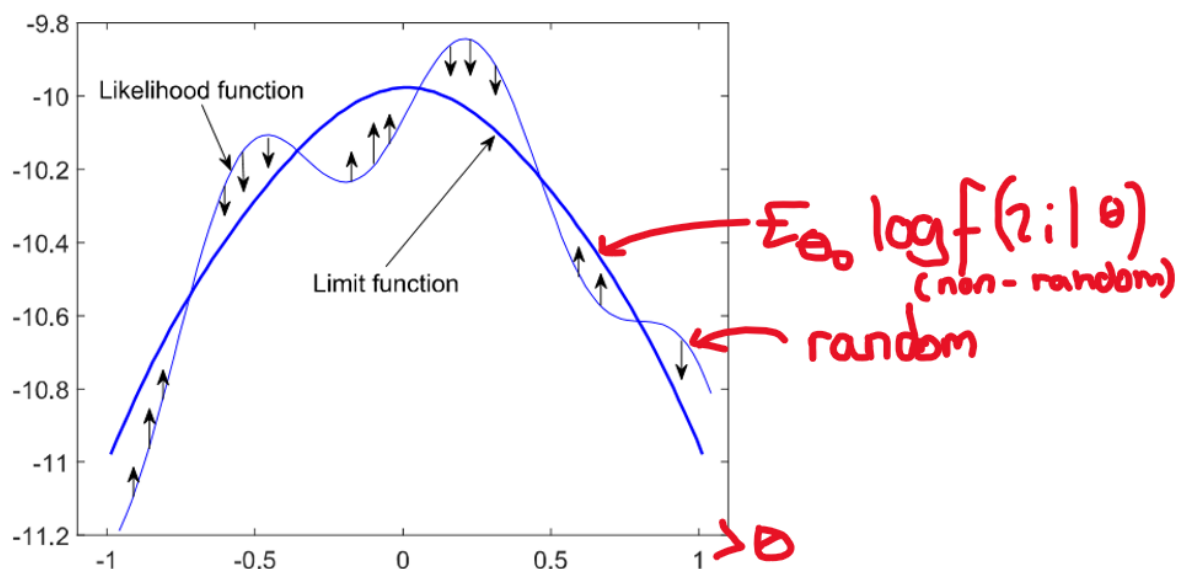
### Proof.

$$\hat{\theta}_{ML} = \operatorname{argmax}_{\theta} \sum_{i=1}^n \log f(z_i; \theta) = \operatorname{argmax}_{\theta} \frac{1}{n} \sum_{i=1}^n \log f(z_i; \theta)$$

But:

$$\frac{1}{n} \sum_{i=1}^n \log f(z_i; \theta) \xrightarrow{p} \mathbb{E}_{\theta_0}[\log f(z; \theta)] \text{ uniformly } \Rightarrow \operatorname{argmax}_{\theta} \frac{1}{n} \sum_{i=1}^n \log f(z_i; \theta) \xrightarrow{p} \operatorname{argmax}_{\theta} \mathbb{E}_{\theta_0}[\log f(z; \theta)]$$

□



**Proposition 11.0.2.**  $E_{\theta_0} \log f(z; \theta)$  is maximised at the true value of parameter  $\theta_0$

**Proof.** Consider the KL divergence between  $f(z; \theta)$  and  $f(z; \theta_0)$ :

$$E_{\theta_0} \log \frac{f(z; \theta_0)}{f(z; \theta)}$$

By construction the minimiser of the KL divergence must be the maximiser of  $E_{\theta_0} \log f(z; \theta)$ .  
By Jensen's inequality:

$$= -E_{\theta_0} \frac{\log f(z; \theta)}{\log f(z; \theta_0)} \geq -\log E_{\theta_0} \frac{f(z; \theta)}{f(z; \theta_0)} = -\log \int \frac{f(z; \theta)}{f(z; \theta_0)} f(z; \theta_0) dz = -\log 1 = 0$$

But we can achieve this bound by setting  $\theta = \theta_0$  is a maximiser of  $E_{\theta_0} \log f(z; \theta)$ .  $\square$

**Note:-**

If there exists another maximiser  $\theta_1$ , we must have  $f(z; \theta_0) = f(z; \theta_1)$  for all  $z$ . In such a case, we say that a case, we say that the parameter is non-identified.

In the linear regression example,  $\theta = (\beta', \sigma^2)$ , would not be identified if  $X'X$  has rank lower than  $k$  (perfect multicollinearity).

Pointwise convergence is not enough for consistency of the  $\theta_{ML}$  estimator. Sufficient conditions are given by uniform convergence and "enough" curvature of  $E_{\theta_0} \log f(z; \theta)$  around  $\theta_0$ .

## 11.0.2 Asymptotic Normality of ML

**Proposition 11.0.3.**

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

where  $I(\theta_0)$  is the Fisher information matrix:

$$I_1(\theta_0) = \text{Var} \left[ \frac{\partial}{\partial \theta} \log f(z; \theta_0) \right] = -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \log f(z; \theta_0) \right] = E_{\theta_0}(H_1) = \mathbb{E}_{\theta_0} \left( \frac{H}{n} \right)$$

Note  $I_1(\theta_0)$  is the Fisher information for a single observation.  
Define  $I(\theta_0)$  as the Fisher information matrix for the sample.

This is the sum of the Fisher information for each observation  $I(\theta_0) = nI_1(\theta_0)$ , since  $\log(z_i; \theta)$  is a function of iid  $z_i$ , and so is iid.

$$\text{Var} \left[ \frac{\partial}{\partial \theta} L(\theta_0) \right] = \text{Var} \left[ \frac{\partial}{\partial \theta} \sum_{i=1}^n \log f(z_i; \theta_0) \right] = n \text{Var} \left[ \frac{\partial}{\partial \theta} \log f(z; \theta_0) \right] \text{ since iid}$$

**Proof.** Let  $\Psi(\theta) = \frac{\partial}{\partial \theta} \frac{1}{n} L(\theta; Z)$ , where

$$L(\theta; Z) = \sum_{i=1}^n \log f(z_i; \theta)$$

$\hat{\theta}_{ML}$  can be obtained as a solution to the likelihood equation:  $\Psi(\hat{\theta}_{ML}) = \frac{\partial}{\partial \theta} \frac{1}{n} L(\hat{\theta}_{ML}; Z) = 0$   
Assuming consistency,  $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$ , it makes sense to expand  $\Psi(\hat{\theta}_{ML})$  around  $\theta_0$ :

$$\Psi(\hat{\theta}_{ML}) = 0 = \Psi(\theta_0) + (\hat{\theta}_{ML} - \theta_0)\Psi'(\theta_0) + \frac{1}{2}(\hat{\theta}_{ML} - \theta_0)^2\Psi''(\tilde{\theta})$$

where  $\tilde{\theta}$  is between  $\hat{\theta}_{ML}$  and  $\theta_0$ , such that the Taylor expansion is exact by the MVT.  
Therefore when  $\theta$  is scalar,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{-\sqrt{n}\Psi(\theta_0)}{\Psi'(\theta_0) + (\hat{\theta}_{ML} - \theta_0)\Psi''(\tilde{\theta})/2}$$

But under the random sampling assumption:

$$\Psi(\theta_0) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(z_i; \theta_0) \Big|_{\theta=\theta_0} \xrightarrow{p} \frac{\partial}{\partial \theta} \mathbb{E}_{\theta_0} \log f(z; \theta_0) \Big|_{\theta=\theta_0}$$

And with the Lindeberg-Levy CLT:

$$-\sqrt{n}\Psi(\theta_0) \xrightarrow{d} N \left( 0, \text{Var} \left( \frac{\partial}{\partial \theta} \log f(z_i, \theta_0) \right) \right) = N \left( 0, \frac{1}{n} I(\theta_0) \right)$$

Next, by Khinchine's LLN:

$$\Psi'(\theta_0) \xrightarrow{p} \frac{\partial^2}{\partial \theta^2} \mathbb{E}_{\theta_0} \log f(z; \theta_0) \Big|_{\theta=\theta_0}$$

Finally,  $(\hat{\theta}_{ML} - \theta_0)\Psi''(\tilde{\theta}) \xrightarrow{p} 0$  i.e. is  $o_p(1)$ , since  $\hat{\theta}_{ML} - \theta_0 = o_p(1)$  and  $\Psi''(\tilde{\theta})$  converges to a finite constant (Amemiya 1985, p. 67, ch 4).

Therefore by Slutsky's theorem:

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{ML} - \theta_0) &= \frac{-\sqrt{n}\Psi(\theta_0)}{\Psi'(\theta_0) + (\hat{\theta}_{ML} - \theta_0)\Psi''(\tilde{\theta})/2} \xrightarrow{d} \frac{N(0, \frac{1}{n}I(\theta_0))}{\mathbb{E}_{\theta_0} \frac{\partial^2}{\partial \theta^2} \log f(z; \theta_0)} \\ \therefore \sqrt{n}(\hat{\theta}_{ML} - \theta_0) &\xrightarrow{d} \frac{N(0, \frac{1}{n}I(\theta_0))}{-\frac{1}{n}I(\theta_0)} = N(0, nI^{-1}(\theta_0)) \end{aligned}$$

In other words in large samples,  $\hat{\theta}_{ML}$  is approximately normally distributed with mean  $\theta_0$  and variance  $I^{-1}(\theta_0)$ .  $\square$

This generalises straightforwardly to the case of a vector  $\theta$ .

NOTE:  $I(\theta_0)$  refers to the sample Fisher information matrix, which is  $n \times I_1(\theta)$  - the finite infor-

mation matrix of one observation. Thus saying  $\hat{\theta}_{ML}$  is approximately normally distributed with mean  $\theta_0$  and variance  $I^{-1}(\theta_0)$ , means its variance is in fact  $(1/n)I_1^{-1}(\theta_0)$ , which goes to zero for large  $n$  and thus we have  $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$  as we found earlier.

## 11.1 Asymptotic efficiency of the maximum likelihood estimator

**Proposition 11.1.1.**  $\theta_{ML}$  is asymptotically efficient:

Lowest asymptotic variance among all estimators that are

- asymptotically normal
- asymptotically unbiased
- regular

Recall the Cramér-Rao result:

Any unbiased estimator of  $\theta_0$  has variance no smaller than the inverse of the Fisher information. While suggestive of asymptotic efficiency here, it is a *finite* sample result and thus does not imply this.

### 11.1.1 Irregular Estimators

Hodges' Estimator

$$\theta_H = \begin{cases} \hat{\theta}_{ML} & \text{if } |\hat{\theta}_{ML}| \geq n^{-1/4} \\ 0 & \text{if } |\hat{\theta}_{ML}| < n^{-1/4} \end{cases}$$

Case 1:  $\theta_0 \neq 0$

$\hat{\theta}_H$  is asymptotically equivalent to  $\hat{\theta}_{ML}$ . This is because  $\hat{\theta}_{ML} \xrightarrow{p} \theta_0 \neq 0$ , and  $n^{-1/4} \rightarrow 0$ , thus  $|\hat{\theta}_{ML}| \geq n^{-1/4}$  will be true asymptotically, so  $\hat{\theta}_H = \hat{\theta}_{ML}$  asymptotically.

**Note:-**

**Big O, Little O Notation**

$f(x) \in O(g(x))$  if  $\exists K > 0$  and  $x_0$  such that  $|f(x)| \leq Kg(x)$  for all  $x > x_0$ .

$f(x) \in o(g(x))$  if  $\forall K > 0 \exists x_0$  such that  $|f(x)| < Kg(x)$  for all  $x > x_0$ .

Product Rule:  $f(x) = O(g(x))$  and  $h(x) = O(k(x)) \Rightarrow f(x)h(x) = O(g(x)k(x))$

Little O  $\Rightarrow$  Big O:  $f(x) = o(g(x)) \Rightarrow f(x) = O(g(x))$

In probability:

$X_n \in O_P(\alpha_n)$  if  $\forall \varepsilon > 0 \exists K > 0$  and  $x_0$  such that  $\Pr(|f(x)| \leq Kg(x)) > 1 - \varepsilon$  for all  $x > x_0$ .

i.e.  $X_n/\alpha_n$  is bounded up to an exceptional event of arbitrarily small (but fixed) positive probability, i.e. the ratio is 'bounded in probability'.

$f(x) \in o_p(\alpha_n)$  if  $\forall \varepsilon > 0 \forall K > 0 \exists x_0$  such that  $\Pr(|f(x)| < Kg(x)) > 1 - \varepsilon$  for all  $x > x_0$ .

Case 2:  $\theta_0 = 0$

**Proposition 11.1.2.**  $|\hat{\theta}_{ML}| = O_p(n^{-1/2})$

Since  $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, I_1^{-1}(\theta_0))$ , we know  $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \in O_p(1)$ , since its variance (and expectation) is finite and constant wrt  $n$  and so must be bounded in probability.

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{\hat{\theta}_{ML} - \theta_0}{1/\sqrt{n}} = O_p(1)$$

$$\Rightarrow \hat{\theta}_{ML} - \theta_0 = O_p(n^{-1/2})^*$$

$$\therefore |\hat{\theta}_{ML}| = O_p(n^{-1/2})$$

\*(also loose intuition from the product rule of normal big O,  $\sqrt{n} = O_p(\sqrt{n})$ )

Where let  $\hat{\theta}_{ML} - \theta_0 \in O_p(\alpha_n)$

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \in O_p(1) \Rightarrow O_P(\sqrt{n})O_P(\alpha_n) = O_P(\sqrt{n}\alpha_n) = O_P(1)$$

$$\Rightarrow \alpha_n = 1/\sqrt{n}$$

**Proposition 11.1.3.**  $|\hat{\theta}_{ML}| = o_p(n^{-1/4})$

$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, I_1^{-1}(\theta_0))$  and  $n^{-1/4} \xrightarrow{p} 0$  Thus by Slutsky's theorem:  $n^{1/4}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} 0$

$$\Rightarrow n^{1/4}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{p} 0$$

$$\Rightarrow \frac{(\hat{\theta}_{ML} - \theta_0)}{1/n^{1/4}} \xrightarrow{p} 0$$

$$\Rightarrow \lim_{n \rightarrow \infty} \mathbb{P}(|\frac{(\hat{\theta}_{ML} - \theta_0)}{1/n^{1/4}} - 0| > \varepsilon) = 0 \forall \varepsilon > 0$$

$$\Rightarrow \hat{\theta}_{ML} - \theta_0 = o_p(n^{-1/4}) \text{ with the definition of } o_p$$

Intuitively as  $n^{-1/4} > n^{-1/2}$ , it makes sense that dividing by  $n^{-1/4}$  binds more strictly (sends to zero) than dividing by  $n^{-1/2}$ , which already binds in probability (sends to a constant variance distribution).

When  $\theta_0 = 0$  Hodges' estimator clearly improves over  $\hat{\theta}_{ML}$  because  $|\hat{\theta}_{ML}| = o_p(n^{-1/4})$ , which implies  $\hat{\theta}_H = 0$  exactly asymptotically (with zero variance) for sufficiently large  $n$ .

But in finite samples, Hodge's estimator behaves poorly for  $\theta \approx 0$ . Asymptotically, this is reflected in its erratic behaviour when true value of parameter is drifting towards zero so that  $\theta = h/\sqrt{n}$  for some  $h \in \mathbb{R}$ . For such sequences of  $\theta$ ,  $\hat{\theta}_H$  is inconsistent. we have:

$$\sqrt{n}(\hat{\theta}_H - \theta_0) = \sqrt{n}(\hat{\theta}_H - h/\sqrt{n}) \rightarrow -h$$

*Regular* estimators would have the same asymptotic distribution for any value of  $h/\sqrt{n}$  (a small change in parameter should not change the distribution of the estimator too much)

## 11.2 Likelihood Ratio Test

Suppose that the likelihood function is in general given by  $L(\theta; Z) \equiv f(Z, \theta)$ , where  $Z$  is a vector of data and  $\theta$  is a vector of parameters. Consider testing the null hypothesis  $H_0 : \theta \in \Theta_0$  against the alternative  $H_1 : \theta \in \Theta_1$ , where  $\Theta_0 \cap \Theta_1 = \emptyset$ .

The likelihood ratio test is defined by the following procedure:

Reject  $H_0$  if

$$LR(Z) = \frac{\sup_{\theta \in \Theta_0} L(\theta; Z)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta; Z)} > c$$

. where  $c$  is chosen as a critical value so as to satisfy  $\max_{\theta \in \Theta_0} \Pr(LR(Z) > c) = \alpha$ , where  $\alpha$  is the significance level of the test (probability of Type 1 error).

**Theorem 11.2.1. Neyman-Pearson Lemma:**

When  $\Theta_0 = \theta_0$  and  $\Theta_1 = \theta_1$  (i.e. single values of the parameter vector), the likelihood ratio test is the most powerful test of size  $\alpha$ .

### 11.2.1 Likelihood Ratio Test of linear restrictions in normal regression

**Proposition 11.2.1.** We show the LR test to be equivalent to the F test, as the LR statistic is a monotone transformation of the F statistic.

Consider a hypothesis  $R\beta = r$  about coefficients of linear regression with normal errors:

$$Y = X\beta + \varepsilon, \varepsilon|X \sim N(0, \sigma^2 I)$$

The unconstrained ML estimates of  $\beta$  and  $\sigma^2$  are in such a model  $\hat{\beta}_{OLS}$  and  $\hat{\sigma}_{ML}^2 = RSS_u/n$ .

We have  $\log(\max_{\theta} L(Y, \theta|X))$  (unrestricted)

$$\begin{aligned} &= \log \left[ \left( \frac{1}{\sqrt{2\pi}|\sigma^2 I|^{-1/2}} \right)^n \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right) \right] \Big|_{\theta=\hat{\theta}_{ML}} \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}_{ML}^2) - \frac{1}{2\hat{\sigma}_{ML}^2} (Y - X\hat{\beta}_{ML})'(Y - X\hat{\beta}_{ML}) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log\left(\frac{RSS_u}{n}\right) - \frac{1}{2} \frac{RSS_u}{RSS_u/n} \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(RSS_u) - \frac{n}{2} \end{aligned}$$

Similarly under the restrictions we can show that:

$$\log(\max_{\theta \in \Theta_0} L(Y, \theta|X)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(RSS_r) - \frac{n}{2}$$

where  $RSS_r$  is the restricted residual sum of squares.

Therefore the log likelihood ratio statistic for the test of  $R\beta = r$  against  $R\beta \neq r$  is:

$$\begin{aligned} LR &= -2 \left[ -\frac{n}{2} \log\left(\frac{RSS_r}{n}\right) + \frac{n}{2} \log\left(\frac{RSS_u}{n}\right) \right] = n \log\left(\frac{RSS_r}{RSS_u}\right) \\ &= n \left[ \log \left( \frac{p}{n-k} \frac{(RSS - r - RSS_u)/p}{RSS_u/(n-k)} + 1 \right) \right] \\ &= n \left[ \log \left( \frac{p}{n-k} \frac{W}{p} + 1 \right) \right] \end{aligned}$$

Thus LR statistic is a monotone transformation of the F statistic  $= W/p$  so that LR test and F test must be equivalent in the context of testing the linear restrictions in normal regression model. But unlike F test, LR test provides a formidable tool for testing hypotheses in much broader contexts.

Finding c:

$$\begin{aligned} P(LR > c) &= P\left(n \log\left(1 + \frac{p}{n-k} F\right) > c\right) \\ &= P\left(F > \frac{n-k}{p} (e^{c/n} - 1)\right) = \alpha \end{aligned}$$

Thus as we know the F distribution:

$$\begin{aligned} \frac{n-k}{p} (e^{c/n} - 1) &= F_{1-\alpha}(p, n-k) \\ \Rightarrow c &= n \log(F_{1-\alpha}(p, n-k) \frac{p}{n-k} + 1) \end{aligned}$$