

## *Maximum Likelihood Estimation*

- Likelihood function and the ML principle;
- Properties of ML estimators;
- Proof of the asymptotic normality
- Estimators of the information matrix
- Regressors
- Robust covariance matrix estimation;
- Hypothesis testing.

# Likelihood function and the ML principle

- Let  $f(y; \theta)$  denote the *probability density function/probability function* of the random variable  $y$ , given  $\theta$ . Our objective is to estimate the true parameter vector  $\theta$ .
- **Example:** A Bernoulli Random variable:

$$Y = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1 - \theta \end{cases}$$

where  $\theta \in (0, 1)$ . Hence

$$f(y; \theta) = \mathcal{P}(Y = y | \theta) = \theta^y (1 - \theta)^{1-y}, \quad y = 0, 1$$

- The *joint density* of  $n$  iid observations of  $y$  is

$$f(y_1, \dots, y_n | \theta) = \prod_{i=1}^n f(y_i; \theta).$$

- If  $y$  is a discrete random variable,  $f(y_1, \dots, y_n | \theta)$  gives the probability of observing a particular sample, given  $\theta$ .

# Likelihood function and the ML principle

- Let us now take  $f(y_i; \theta)$  as a function of  $\theta$  given  $y$  and write

$$L(\theta|y_1, \dots, y_n) = \prod_{i=1}^n f(y_i; \theta).$$

- This is the *likelihood function*, which gives the likelihood that the population parameter is  $\theta$ , given the observed sample.
- Note:  $L(\theta|y_1, \dots, y_n)$  is often abbreviated to  $L(\theta)$ .

# Likelihood function and the ML principle

- The *Maximum Likelihood (ML) principle* suggests that estimators of the unknown parameters are obtained by maximizing  $L(\theta)$  with respect to  $\theta$ .

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

It is often convenient to work with the natural logarithm of the likelihood function  $\log L(\theta)$ . For example, in the iid case:

$$\log L(\theta | y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i; \theta);$$

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

# Likelihood function and the ML principle

- Usually  $\hat{\theta}$  can be obtained by solving the *likelihood equation*

$$\left. \frac{\partial \log L(\theta)}{\partial \theta} \right|_{\hat{\theta}} = 0$$

- **Example:** In the Bernoulli case  $\mathcal{P}(Y = y|\theta) = \theta^y(1 - \theta)^{1-y}$  we have

$$\log L(\theta) = \sum_{i=1}^n y_i \log(\theta) + \sum_{i=1}^n (1 - y_i) \log(1 - \theta).$$

the solution is given by  $\hat{\theta} = \bar{y} = \sum_{i=1}^n y_i / n$ .

- Occasionally the ML estimator is *not unique*.
- Also,  $\log L(\theta)$  may have only one global maximum, but multiple *local maxima*.

# Likelihood function and the ML principle

- The main *regularity conditions* (now assumed to hold) are as follows:
  - 1 The first three derivatives of  $\log f(y|\theta)$  with respect to  $\theta$  are continuous and finite for almost all  $y$  and for all  $\theta$ ;
  - 2 For all values of  $\theta$ ,  $\left| \partial^3 \log f(y|\theta) / \partial \theta_j \partial \theta_k \partial \theta_l \right|$  is limited by a function that has finite expectation;
  - 3 The domain of  $y$  does not depend on  $\theta$ ;
  - 4  $\theta$  is an interior point to the compact parameter space  $\Theta$ .

# Likelihood function and the ML principle

- In order to proceed, it is interesting to look at some important results (Bartlett identities).
- Define the *score* vector  $S(\theta)$  and the *Hessian* matrix  $H(\theta)$  as

$$S(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \log f(y_i | \theta)}{\partial \theta} = \sum_{i=1}^n s_i(\theta),$$
$$H(\theta) = \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^n \frac{\partial^2 \log f(y_i | \theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^n H_i(\theta).$$

- *First Bartlett identity*:  $E[s_i(\theta)] = 0$ .
- Hence  $E[S(\theta)] = 0$
- *Second Bartlett identity*:  $\text{Var}[s_i(\theta)] = -E[H_i(\theta)]$
- $\text{Var}[s_i(\theta)] = E[s_i(\theta) s_i(\theta)']$  defines Fisher's *information matrix*, denoted  $\mathcal{I}(\theta)$ .
- Hence, the result  $\text{Var}[s_i(\theta)] = E[s_i(\theta) s_i(\theta)'] = -E[H_i(\theta)]$  is also called the *information matrix identity*.

# Properties of MLE

- Under the assumed regularity conditions the MLE possesses the following properties:
  - 1 **Consistency:**  $\text{plim } \hat{\theta} = \theta$ ;
  - 2 **Asymptotic normality:**  $\sqrt{n} (\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta)^{-1})$ ;
  - 3 **Asymptotic efficiency:** if  $\tilde{\theta}$  is a regular consistent asymptotically normal estimator such that  $\sqrt{n} (\tilde{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Omega)$ , then  $\Omega - [\mathcal{I}(\theta)]^{-1}$  is positive semi-definite, i.e., under these RC, the MLE asymptotically achieves the **Cramer-Rao** lower bound which is given by  $[\mathcal{I}(\theta)]^{-1}$ ;
  - 4 **Invariance:** If  $c(\theta)$  is a continuous and continuously differentiable one-to-one function, the MLE of  $\gamma = c(\theta)$  is  $c(\hat{\theta})$ .
- **Example:** in the Bernoulli case  $\mathcal{I}(\theta)^{-1} = \theta(1 - \theta)$ , therefore  $\sqrt{n} (\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta(1 - \theta))$ ;



# Proof of the asymptotic normality

- It is enlightening to sketch the proof of the asymptotic normality.
- Under the assumed regularity conditions, we have

$$\begin{aligned}\frac{\partial \log L(\hat{\theta})}{\partial \theta} &= 0, \\ \sum_{i=1}^n \frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta} &= 0\end{aligned}$$

- Expanding this result in a *1st order Taylor series* around  $\theta$  we have

$$\begin{aligned}\sum_{i=1}^n \frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta} &= \sum_{i=1}^n \frac{\partial \log f(y_i|\theta)}{\partial \theta} + \sum_{i=1}^n \frac{\partial^2 \log f(y_i|\bar{\theta})}{\partial \theta \partial \theta'} (\hat{\theta} - \theta) = 0 \\ \underbrace{\sqrt{n} \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f(y_i|\theta)}{\partial \theta}}_{\hat{S}(\theta)} &+ \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(y_i|\bar{\theta})}{\partial \theta \partial \theta'}}_{\hat{H}(\bar{\theta})} \sqrt{n} (\hat{\theta} - \theta) = 0\end{aligned}$$

where  $\bar{\theta} = w\hat{\theta} + (1-w)\theta$  for  $0 \leq w \leq 1$ .

# Proof of the asymptotic normality

- Write

$$\sqrt{n}(\hat{\theta} - \theta) = [-\hat{H}(\bar{\theta})]^{-1} \sqrt{n}\hat{S}(\theta)$$

Notice that, because  $\text{plim}(\hat{\theta} - \theta) = 0$ , and we have  $\text{plim}(\bar{\theta} - \theta) = 0$  and (under some conditions)

$$\text{plim} -\hat{H}(\bar{\theta}) = \text{E}[-H_i(\theta)] = A$$

- Now we can apply a Central Limit Theorem for random samples

to obtain  $\sqrt{n}\hat{S}(\theta) \xrightarrow{d} \mathcal{N}(0, B)$  where

$$B = \text{Var}[s_i(\theta)] = \text{E}[s_i(\theta)s_i(\theta)']$$

- Recall that if  $x \sim \mathcal{N}(0, C)$  then  $Dx \sim \mathcal{N}(0, DCD')$ .

- Hence

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1})$$

$$A = \text{E}[-H_i(\theta)] \quad B = \text{E}[s_i(\theta)s_i(\theta)']$$

- For correctly specified models,  $B = \text{E}[-H_i(\theta)] = A = \mathcal{I}(\theta)$  and

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta)^{-1});$$

# Estimators of the information matrix

- There are three commonly used estimators of  $\mathcal{I}(\theta)$ 
  - 1 *Expected Information*: If the form of the expected values of the second derivatives of the log-likelihood function is known, then we can evaluate  $\mathcal{I}(\theta)$  at  $\hat{\theta}$ .
  - 2 *Observed Information*: Simply use  $-\hat{H}(\hat{\theta})$ .
  - 3 *Outer Product of the Gradient (OPG)*: Because of the information matrix identity, we can also use  $n^{-1} \sum_{i=1}^n s_i(\hat{\theta}) s_i(\hat{\theta})'$ .
- The *OPG* is notorious for its poor finite sample performance.

- The previous results are easy to extend to accommodate the presence of covariates.
- Suppose the joint distribution of  $y$  and  $x$  depends on  $\alpha$ , giving  $f(y, x|\alpha) = f(y|x, \alpha)g(x|\alpha)$ .
- Next, suppose that  $\alpha$  can be divided into  $\theta$  and  $\delta$ , so that (exogeneity of  $x$ )  $f(y, x|\alpha) = f(y_i|x_i, \theta)g(x_i|\delta)$ .
- For an iid sample  $\{(y_i, x_i)\}_{i=1}^n$  then

$$\log L(\theta, \delta|y_i, x_i) = \sum_{i=1}^n \log f(y_i, x_i|\alpha) = \sum_{i=1}^n \log f(y_i|x_i, \theta) + \sum_{i=1}^n \log g(x_i|\delta).$$

and the term  $\sum_{i=1}^n \log g(x_i|\delta)$  can be ignored

- $\hat{\theta}$  can then be obtained by maximizing just  $\sum_{i=1}^n \log f(y_i|x_i, \theta)$  with respect to  $\theta$ . Therefore, frequently we will work directly with the conditional log-likelihood

$$\log L(\theta|y_i, x_i) = \sum_{i=1}^n \log f(y_i|x_i, \theta),$$

and this (under appropriate regularity conditions) will behave to a large extent like a standard log-likelihood.

- However, now  $E[-H_i(\theta)] = E\left[-\frac{\partial^2 \log f(y_i|x_i, \theta)}{\partial \theta \partial \theta'}\right] = \mathcal{I}(\theta) = E\left[\frac{\partial \log f(y|x_i, \theta)}{\partial \theta} \frac{\partial \log f(y|x_i, \theta)}{\partial \theta'}\right]$ , and so on.

# Robust covariance matrix estimation

- If the likelihood function is misspecified, the MLE is generally inconsistent for the parameters of interest.
- However, under very general conditions,  $\text{plim } \hat{\theta} = \theta^*$ , where the *pseudo-true value*  $\theta^*$  minimizes the Kullback-Leibler divergence, that is

$$\theta^* = \arg \min_c \int_{-\infty}^{+\infty} \left[ \log \left( \frac{f_0(y)}{f(y|c)} \right) \right] f_0(y) dy = \arg \min_c E \left[ \log \left( \frac{f_0(y)}{f(y|c)} \right) \right].$$

where  $f_0(y)$  is the true distribution of the data.

- The *Kullback–Leibler divergence* (also called relative entropy) is a measure of how one probability distribution is different from a second, reference probability distribution
- That is, the MLE leads to the *best approximation*, in the Kullback-Leibler sense, to  $f_0(y)$ , the true density.
- However, because the IM identity does not hold, the asymptotic covariance matrix is given by:

$$A^{-1}BA^{-1}, \quad A = E[-H_i(\theta^*)] \quad B = E[g_i(\theta^*)g_i(\theta^*)'].$$

# Hypothesis testing

- Consider a general set of restrictions to be tested

$$H_0 : h(\theta) = 0$$

where:

- $\theta$ : vector of parameters in model
- $h(\theta)$ :  $d \times 1$  vector of restrictions
- $L(\theta)$ : likelihood function for model
- $S(\theta) = \sum_{i=1}^n s_i(\theta)$  the efficient score
- $\hat{\theta}$  and  $\tilde{\theta}$ : unrestricted and restricted MLE, respectively (that is  $\hat{\theta} = \hat{\theta}_{ml}$  and  $\tilde{\theta} = \tilde{\theta}_{ml}$ ).
- $\hat{\theta}$  is the value of  $\theta$  that maximizes  $\log L(\theta)$
- $\tilde{\theta}$  is the value of  $\theta$  that maximizes  $\log L(\theta)$  and satisfy  $h(\theta) = 0$ .
- $L(\hat{\theta})$  and  $L(\tilde{\theta})$ : value of  $L(\theta)$  evaluated at  $\hat{\theta}$  and  $\tilde{\theta}$ , respectively.

# The 3 classical test principles

## *Likelihood Ratio Tests:*

- Compare  $L(\hat{\theta})$  and  $L(\tilde{\theta})$  (if  $h(\theta) = 0$  then  $L(\hat{\theta})$  should be close to  $L(\tilde{\theta})$ )

## *Wald Tests:*

- Compare  $h(\hat{\theta})$  with 0 (since  $h(\tilde{\theta}) = 0$ ).

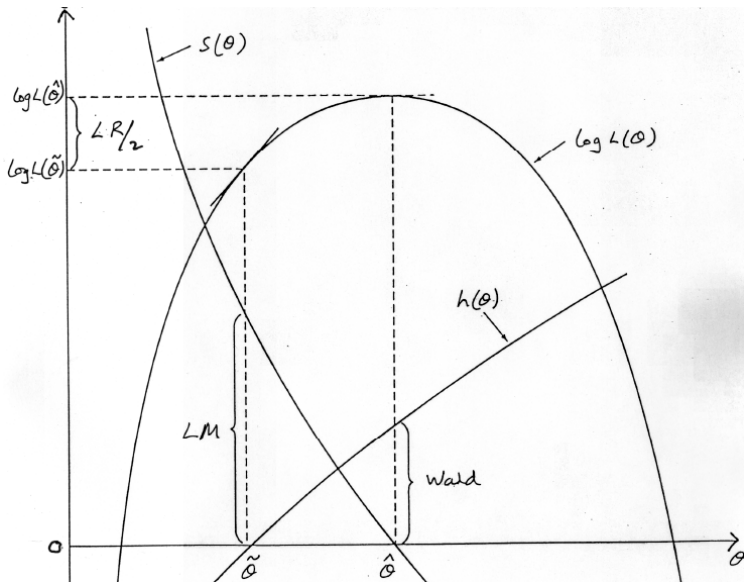
## *Lagrange Multiplier or Score Tests:*

- Compare  $S(\tilde{\theta})$  with 0 (since  $S(\hat{\theta}) = 0$ ).



# The 3 classical test principles

Intuition:



# The Wald Test

- How close is  $h(\hat{\theta})$  to zero (since  $h(\tilde{\theta}) = 0$ )?
- Test statistic:

$$\mathcal{W} = n \times h(\hat{\theta})' \left[ G(\hat{\theta})' \left[ \widehat{\mathcal{I}(\theta)} \right]^{-1} G(\hat{\theta}) \right]^{-1} h(\hat{\theta}).$$

where

$$G(\theta) = \frac{\partial h(\theta)}{\partial \theta}.$$

and  $\widehat{\mathcal{I}(\theta)}$  is an estimator of  $\mathcal{I}(\theta)$ .

- Under the null hypothesis:

$$\mathcal{W} \xrightarrow{D} \chi^2(d).$$

**Shortcoming:** Wald test not invariant to how restrictions are formulated. E.g.:  $\beta / (1 - \alpha) = 1$  (nonlinear restriction) and  $\beta + \alpha - 1 = 0$  (linear restriction) are equivalent restrictions, but may lead to different values of  $\mathcal{W}$ .

**Note:** If  $h(\theta) = R\theta - q$ , and  $G(\theta) = R$ .

# The Likelihood Ratio Test

- How “close” are  $\mathcal{L}(\hat{\theta})$  and  $\mathcal{L}(\tilde{\theta})$ ?
- Test is based on the *likelihood ratio*:

$$\lambda = \frac{\mathcal{L}(\tilde{\theta})}{\mathcal{L}(\hat{\theta})}.$$

- Test statistic

$$\begin{aligned}\mathcal{LR} &= -2 \log(\lambda) \\ &= 2\{\log \mathcal{L}(\hat{\theta}) - \log \mathcal{L}(\tilde{\theta})\}\end{aligned}$$

- Under the null hypothesis:

$$\mathcal{LR} \xrightarrow{D} \chi^2(d)$$

# The Lagrange Multiplier (LM) or Score Test

- How close is  $S(\tilde{\theta})$  to zero (since  $S(\hat{\theta}) = 0$ )?
- Test statistic

$$\mathcal{LM} = S(\tilde{\theta})' \left[ \widehat{\mathcal{I}(\theta)} \right]^{-1} S(\tilde{\theta}) / n$$

- Under the null hypothesis:

$$\mathcal{LM} \xrightarrow{D} \chi^2(d).$$