Maximum Likelihood Estimation

- Likelihood function and the ML principle;
- Properties of ML estimators;
- Estimators of the information matrix
- Regressors
- Interpretation and inference in misspecified models
- Hypothesis testing.

Likelihood function and the ML principle

- Let $f(y;\theta_0)$ denote the *probability density function/probability function* of the random variable y, given θ_0. Our objective to to estimate the true parameter vector θ_0.
- **Example:** A Bernoulli Random variable:

$$Y = \begin{cases} 1 & \text{with probability } \theta_0 \\ 0 & \text{with probability } 1-\theta_0 \end{cases}$$

where $\theta_0 \in (0,1)$.Hence

$$f(y;\theta_0) = \mathcal{P}(Y = y|\theta_0) = \theta_0^y(1-\theta_0)^{1-y},\ y = 0,1$$

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

$$f(y_1,\ldots,y_n|\theta_0) = \prod_{i=1}^{n} f(y_i;\theta_0).$$

- If y is a discrete random variable, $f(y_1,\ldots,y_n|\theta_0)$ gives the probability of observing a particular sample, given θ_0.

Likelihood function and the ML principle

- Let us now take $f(y_i;\theta)$ as a function of θ given $y_1, \ldots, y_n$ and write

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

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

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

- If the random variable y is discrete $\hat{\theta}$ yields the value of θ that maximizes the probability of observing the sample that actually occurred. If the random variable y is continuous, such interpretation is no longer valid.

Likelihood function and the ML principle

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

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

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

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 $\mathcal{P}(Y = y|\theta_0) = \theta_0^y(1-\theta_0)^{1-y}$ 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 θ are continuous and finite for almost all y and for all θ;
  2. For all values of θ, $\left|\partial^3 \log f(y|\theta) / \partial\theta_j \partial\theta_k \partial\theta_l\right|$ is limited by a function that has finite expectation;
  3. The domain of y does not depend on θ;
  4. θ is an interior point to the compact parameter space Θ.

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

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

- *First Bartlett identity*: $\mathrm{E}[s_i(\theta_0)] = 0$.
- Hence $\mathrm{E}[S(\theta_0)] = 0$
- *Second Bartlett identity:* $\mathrm{Var}[s_i(\theta_0)] = -\mathrm{E}[H_i(\theta_0)]$
- $\mathrm{Var}[s_i(\theta_0)] = \mathrm{E}\left[s_i(\theta_0)\, s_i(\theta_0)'\right]$ defines Fisher's *information matrix*, denoted $\mathcal{I}(\theta_0)$.
- Hence, the result $\mathrm{Var}[s_i(\theta_0)] = \mathrm{E}\left[s_i(\theta_0)\, s_i(\theta_0)'\right] = -\mathrm{E}[H_i(\theta_0)]$ is also called the *information matrix identity*.

Properties of MLE

- Under the assumed regularity conditions the MLE possesses the following properties:
  1. *Consistency*: plim $\hat{\theta} = \theta_0$;
  2. *Asymptotic normality*: $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$;
  3. *Asymptotic efficiency*: if $\tilde{\theta}$ is a regular consistent asymptotically normal estimator such that $\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \Omega)$, then $\Omega - [\mathcal{I}(\theta_0)]^{-1}$ is positive semi-definite, i.e., under these RC, the MLE asymptotically achieves the *Cramer-Rao* lower bound which is given by $[\mathcal{I}(\theta_0)]^{-1}$;
  4. *Invariance:* If $c\left(\theta\right)$ is a continuous and continuously differentiable one-to-one function, the MLE of $\gamma = c(\theta_0)$ is $c\left(\hat{\theta}\right)$.
- **Example:** in the Bernoulli case $\mathcal{I}(\theta_0)^{-1} = \theta_0(1-\theta_0)$, therefore $\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0, \theta_0(1-\theta_0))$;

Estimators of the information matrix

- There are three commonly used estimators of $\mathcal{I}(\theta_0)$
  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 $-\hat{H}\left(\hat{\theta}\right)$.
  3. *Outer Product of the Gradient (OPG)*: Because of the information matrix identity, we can also use $n^{-1}\sum_{i=1}^{n} s_i\left(\hat{\theta}\right) s_i\left(\hat{\theta}\right)'$.
- The *OPG* is notorious for its poor finite sample performance.

Regressors

- The previous results are easy to extend to accommodate the presence of covariates.
- Suppose the joint distribution of y and x depends on α_0, giving $f(y,x|\alpha_0) = f(y|x,\alpha_0)g(x|\alpha_0)$.
- Next, suppose that α_0 can be divided into θ_0 and δ_0, so that (exogeneity of x) $f(y,x|\alpha_0) = f(y_i|x_i,\theta_0)g(x_i|\delta_0)$.
- 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

Regressors

- $\hat{\theta}$ can then be obtained by maximizing just $\sum_{i=1}^{n} \log f(y_i|x_i, \theta)$ with respect to θ. 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 $\mathrm{E}[-H_i(\theta_0)] = E\left[-\frac{\partial^2 \log f(y_i|x_i,\theta_0)}{\partial\theta\partial\theta'}\right] = \mathcal{I}(\theta_0) = \mathrm{E}\left[\frac{\partial \log f(y|x_i,\theta_0)}{\partial\theta}\frac{\partial \log f(y|x_i,\theta_0)}{\partial\theta'}\right]$, and so on.

Interpretation and inference in misspecified models

- 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* θ^* minimizes the Kullback-Leibler divergence, that is

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

  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:

$$A^{-1}BA^{-1}, \quad A = \mathrm{E}\left[-H_i\left(\theta^*\right)\right] \qquad B = \mathrm{E}\left[s_i(\theta^*)s_i(\theta^*)'\right].$$

Hypothesis testing

- Consider a general set of restrictions to be tested

$$H_0 : h(\theta_0) = 0$$

where:

- θ: vector of parameters in model (the true value is given by θ_0)
- $h(\theta)$: $d \times 1$ 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 θ that maximizes $\log L(\theta)$
- $\tilde{\theta}$ is the value of θ 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 $h(\theta_0) = 0$ then $L(\hat{\theta})$ should be close to $L(\tilde{\theta})$

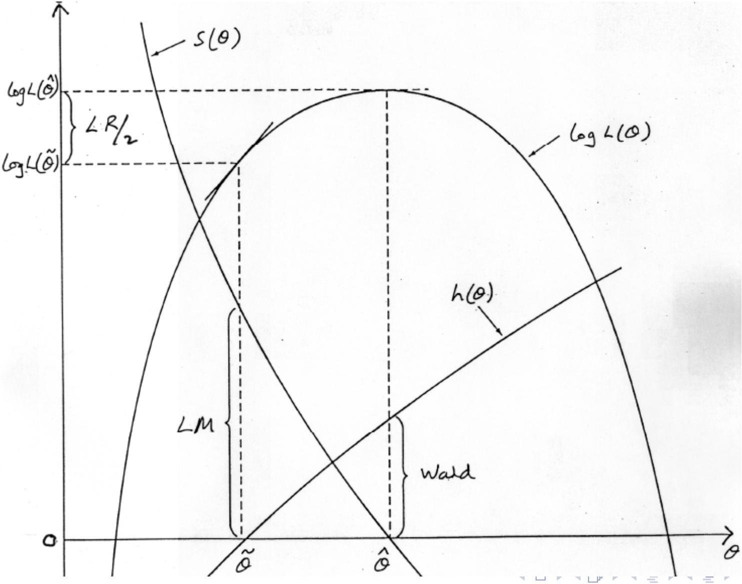
Wald Tests:

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

Lagrange Multiplier or Score Tests:

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

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 $\mathcal{I}(\theta_0)$.
- Under the null hypothesis:

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

The Wald Test

Shortcoming: Wald test not invariant to how restrictions are formulated. E.g.: $\beta_0 / (1 - \alpha_0) = 1$ (nonlinear restriction) and $\beta_0 + \alpha_0 - 1 = 0$ (linear restriction) are equivalent restrictions, but may lead to different values of $\mathcal{W}$.

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 $S(\widetilde{\theta})$ to zero (since $S(\hat{\theta}) = 0$)?
- Test statistic
$$\mathcal{LM} = S(\widetilde{\theta})' \left[\widehat{\mathcal{I}(\theta)}\right]^{-1} S(\widetilde{\theta})/n$$
- Under the null hypothesis:
$$\mathcal{LM} \xrightarrow{D} \chi^2(d).$$