last change 01.11.12@18.40

1.1 Maximum Likelihood Estimation

Example: n Bernoulli trials $0-1$-sequence $\mathbf{y} = (y_1, \ldots, y_n)$. success probability p.

$$L(p \mid \mathbf{y}) = \prod_1^n p^{y_i}(1-p)^{1-y_i}$$

ml estimator

$$l(p \mid \mathbf{y}) = \log L(p \mid y) = \sum y \log p + (n - \sum y) \log(1-p)$$

MLE[1]

$$\widehat{p} = \arg\max_p l(p|\ x) = \frac{\sum y}{n}$$

properties:

- unbiasedness: $\mathbb{E}\widehat{p} = p$
- variance: $\mathbb{V}\widehat{p} = p(1-p)/n$
- consistency: $\mathbb{P}(\ |\widehat{p} - p| > \epsilon) \to 0$ as $n \to \infty$ (law of large numbers)
- asymptotic distribution: (CLT) as $n \to \infty$

$$\frac{\widehat{p} - p}{\sqrt{p(1-p)/\,n}} \to \mathcal{N}(0,1)$$

properties of the mle

density:

$$y_1, \ldots, y_n \text{ iid} \sim f_\theta$$

in the discrete case, replace the density by probability p_θ

likelihood

$$L\,(\,\theta \mid y_1, \ldots, y_n) = f_\theta\,(y_1, \ldots, y_n) = \prod f_\theta\,(y_i) = \prod L\,(\,\theta \mid y_i)$$

take logs for numerical and mathematical (CLT) reasons

[1] **note**: strictly, maximization has to take place under the restrictions $0 \le p \le 1$.By introducing the reparametrization

$$p = \frac{e^\beta}{1+e^\beta}$$

this maximization in β is without restriction. this will be used in the logit model.

$$l(\theta \mid y_1, \ldots, y_n) = \log L(\theta \mid y_1, \ldots, y_n) = \log f_\theta(y_1, \ldots, y_n)$$
$$= \sum \log f_\theta(y_i) = \sum \log L(\theta \mid y_i) = \sum l(\theta \mid y_i)$$

maximize

$$\widehat{\theta} = \arg\max_\theta l(\theta \mid y_1, \ldots, y_n)$$

notes

1. as the likelihood is a function of the unknown parameter given the observations
2. as the observations are random variables the likelihood function in each possible value θ is a random variable of the observations
3. the likelihood in every possible value θ has expectation and variance w.r. to the random observations
4. the same hods for $\widehat{\theta}$

define **Fisher information:**

$$\mathbb{I}_n(\theta) = -\mathbb{E}\left(\sum_i^n \frac{d^2}{d\theta^2} l(\theta \mid y_i)\right)$$

under *iid assumption*

$$\mathbb{I}_n(\theta) = -n\mathbb{E}\left(\frac{d^2}{d\theta^2} l(\theta \mid y)\right) = -n\int \frac{d^2}{d\theta^2} l(\theta \mid y)\ \ f_\theta(y)\, dy = n\mathbb{I}_1$$

Notes

1. the likelihood function is supposed to be **log-concave** in order to have a unique maximum.
2. the (Fisher) information increases linearly with the sample size
3. the (Fisher) information tends to ∞ for i.i.d. observations
4. the Fisher information is a misnomer and should not be confounded with the Shannon information

asymptotic normality of mle:

$$\sqrt{\mathbb{I}_n}\left(\widehat{\theta} - \theta\right) \to \mathcal{N}(0, 1)$$

sometimes this is written loosely as $\widehat{\theta} \to \mathcal{N}\left(\theta, \mathbb{I}_n^{-1}\right)$
Note

$$\mathbb{V}\widehat{\theta} \to \mathbb{I}_n^{-1}$$

Furthermore, it is *asymptotically unbiased*: $\mathbb{E}\widehat{\theta} \to \theta$

Example: for the binomial probability (n Bernoulli trials)

$$l(p \mid \mathbf{y}) = \log L(p \mid \mathbf{y}) = \sum y \log p + (n - \sum y) \log(1-p)$$

$\Longrightarrow$

$$\begin{aligned}
\frac{d}{dp} l(p \mid \mathbf{y}) &= \frac{\sum y}{p} - \frac{n - \sum y}{1-p} \\
\frac{d^2}{dp^2} l(p \mid \mathbf{y}) &= -\frac{\sum y}{p^2} - \frac{n - \sum y}{(1-p)^2} \\
\mathbb{E}\frac{d^2}{dp^2} l(p \mid \mathbf{y}) &= -\left(\frac{np}{p^2} + \frac{n - np}{(1-p)^2}\right) \\
&= -n\left(\frac{1}{p} + \frac{1}{1-p}\right) = \frac{-n}{p(1-p)}
\end{aligned}$$

$\Longrightarrow$

$$\sqrt{\mathbb{I}_n}\,(\widehat{p} - p) = \sqrt{\frac{n}{p(1-p)}}\,(\widehat{p} - p) \to \mathcal{N}(0,1)$$

Notes

1. asymptotic normality $\mathbb{I}(\theta)$ depends on the unknown value replace by $\mathbb{I}\left(\widehat{\theta}\right)$
2. if the expectation required in the computation of $\mathbb{I}(\theta)$ may be hard to find, estimate $\mathbb{I}(\theta)$ by $\widehat{\mathbb{I}}_n = \sum_i^n \frac{d^2}{d\theta^2} l\left(\widehat{\theta} \middle| \ y_i\right)$
3. loosely it could be stated

   $$\widehat{\theta} \to \mathcal{N}\left(\theta, \widehat{\mathbb{I}}_n^{-1}\right)$$

   as $n \to \infty$, where

   $$\widehat{\mathbb{I}}_n = -\sum \frac{d^2}{d\theta^2} l\left(\widehat{\theta} \,\middle|\, y_i\right)$$

higher dimensional parameters

sample $Y_1, \ldots, Y_n$ iid $f(y|\ \theta)$, $\theta' = (\theta_1, \ldots, \theta_J) \implies$ log likelihood function:

$$l(\theta \mid y_1, \ldots, y_n) = l(\theta \mid \mathbf{y}) = \sum\nolimits_i^n \log f(y_i \mid \theta) = \sum\nolimits_i^n l(\theta \mid y_i)$$

mle

$$\widehat{\theta} = \arg\min_{\theta} l(\theta \mid \mathbf{y})$$

score function ($J-$vector):

$$\frac{\partial}{\partial\theta}l(\theta\ |\mathbf{y}) = \left[\frac{\partial}{\partial\theta_j}l(\theta\ |\mathbf{y})\right] = \sum\nolimits_i^n \frac{\partial}{\partial\theta}l(\theta\ |y_i)$$

mle ($J-$vector) solution to score equation:

$$\frac{\partial}{\partial\theta}l(\widehat{\theta}\ |\mathbf{y}) = \left[\frac{\partial}{\partial\theta_j}l(\widehat{\theta}\ |\mathbf{y})\right]_{j=1}^{J} = \mathbf{0}$$

Fisher information ($J \times J-$matrix) of n observations ($\mathbb{I}_n$) resp of 1 observation ($\mathbb{I}_1$).

$$\mathbb{I}_n(\theta) = -\mathbb{E}\frac{\partial^2}{\partial\theta^2}l(\theta\ |\mathbf{y}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_{j_1}\partial\theta_{j_2}}l(\theta\ |\mathbf{y})\right]_{j_1,j_2}^{J,J} = -\left[\sum\nolimits_i^n \mathbb{E}\frac{\partial^2}{\partial\theta_{j_1}\partial\theta_{j_2}}l(\theta\ |y)\right]$$

Notes:

1. In the iid case $\mathbb{I}_n = n\mathbb{I}_1$
2. since $\mathbb{I}_n(\theta)$ is a positive definite symmetric matrix it can be factored into the product of two matrices[2] $\left(\sqrt{\mathbb{I}_n}\right)'\sqrt{\mathbb{I}_n} = \mathbb{I}_n$.

asymtotic normality of mle:

$$\sqrt{\mathbb{I}_n}\left(\widehat{\theta} - \theta\right) \to \mathcal{N}_J(0, I)$$

this holds under certain under regularity conditions (existence of derivatives of log likelihoods etc, *true value parameter not boundary value* etc)

LR-test

$$\begin{aligned} \mathrm{H}_0 &: \ \theta \in \mathbf{\Theta}_0 \\ \mathrm{H}_1 &: \ \theta \in \mathbf{\Theta} - \mathbf{\Theta}_0 \end{aligned}$$

where $\mathbf{\Theta}_0 \subset \mathbf{\Theta}$

compute

$$\begin{aligned} \widehat{\theta}_0 &= \arg\min_{\theta\in\mathbf{\Theta}_0} l(\theta\ |\mathbf{y}) \\ \widehat{\theta} &= \arg\min_{\theta\in\mathbf{\Theta}} l(\theta|\ \mathbf{y}) \end{aligned}$$

[2] this is realted to the **principal axis theorem,** (see wikipedia and elsewhere)

It can be shown that

$$2\sum\nolimits_i^n \left(l(\widehat{\theta}|\ y_i) - l(\widehat{\theta}_0|\ y_i) \right) \rightarrow \chi^2 \left(df \right)$$

where $df = \dim\left(\boldsymbol{\Theta}\right) - \dim\left(\boldsymbol{\Theta}_0\right)$ number of free parameters

Wald test

(case $J = 1$) under $\mathrm{H}_0 : \theta{=}\theta_0$

$$\sqrt{\mathbb{I}_n\left(\theta_0\right)}\left(\widehat{\theta} - \theta_0\right) \rightarrow \mathcal{N}_1\left(0,1\right) \implies \mathbb{I}_n\left(\theta_0\right)\left(\widehat{\theta}{-}\theta_0\right)^2 \rightarrow \chi^2\left(1\right)$$

This follows directly from the asymptotic normality of the ML-estimator (Taylor developing the score function). For case $p \geq 1$

$$\left(\widehat{\theta}{-}\theta_0\right)' \mathbb{I}_n\left(\theta_0\right)\left(\widehat{\theta}{-}\theta_0\right) \rightarrow \chi^2\left(p\right)$$