Maximum likelihood theory (overview)

Suppose we have a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the distribution with a density $f(\mathbf{x}; \boldsymbol{\theta})$ with respect to a σ-finite measure μ and that the density is known up to unknown p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T} \in \Theta$. Let $\boldsymbol{\theta}_X = (\theta_{X1}, \ldots, \theta_{Xp})^\mathsf{T}$ be the true value of the parameter.

Define the *likelihood function* as

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{X}_i; \boldsymbol{\theta})$$

and the *log-likelihood function* as

$$\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

The *maximum likelihood estimator* of parameter $\boldsymbol{\theta}_X$ is defined as

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\, L_n(\boldsymbol{\theta}).$$

Usually we search for the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ as a solution of the system of likelihood equations $\mathbf{U}_n\big(\widehat{\boldsymbol{\theta}}_n\big) \overset{!}{=} \mathbf{0}$, where the random vector

$$\mathbf{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{X}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \log f(\boldsymbol{X}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is called *the score statistic.*

Under appropriate regularity assumptions

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\big) \xrightarrow[n\to\infty]{d} \mathsf{N}_p\big(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_X)\big),$$

where

$$I(\boldsymbol{\theta}_X) = -\mathsf{E}\, \frac{\partial^2 \log f(\boldsymbol{X}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}}\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_X}$$

is the Fisher information matrix.

To make inference about $\boldsymbol{\theta}_X$ usually one needs to estimate the information matrix $I(\boldsymbol{\theta}_X)$. In regression context we usually use *the observed information matrix* defined at $\widehat{\boldsymbol{\theta}}_n$ which is defined as

$$\widehat{I}_n = -\frac{1}{n} \frac{\partial \mathbf{U}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(\boldsymbol{X}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}}\Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n}.$$

Inference about the vector parameter $\boldsymbol{\theta}$

Suppose we want to test the null hypothesis $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ against the alternative $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$. One of the possible test is *Wald test* and it is based on the following test statistic

$$W_n = n\,\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)^\mathsf{T}\, \widehat{I}_n\, \big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big).$$

It can be shown that under the null hypothesis W_n converges in distribution to a χ^2-distribution with p degrees of freedom.

The (asymptotic) confidence set for $\boldsymbol{\theta}_X$ is then constructed as

$$\left\{\boldsymbol{\theta};\, n\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)^{\mathsf{T}} \widehat{I}_n \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \leq \chi_p^2(1-\alpha)\right\},$$

where $\chi_p^2(1-\alpha)$ is the $1-\alpha$ quantile of χ^2-distribution with p degrees of freedom.

Inference about θ_{Xk} (the k-th coordinate of $\boldsymbol{\theta}_X$)

Suppose we want to test the null hypothesis $H_0 : \theta_{Xk} = \theta_0$ against the alternative $H_1 : \theta_{Xk} \neq \theta_0$. One of the possible test is *Wald test* and it is based on the following test statistic

$$T_n = \frac{\sqrt{n}\left(\widehat{\theta}_{nk} - \theta_0\right)}{\sqrt{i_n^{kk}}},$$

where $\widehat{\theta}_{nk}$ is the k-th element of $\widehat{\boldsymbol{\theta}}_n$ and i_n^{kk} is the k-th diagonal element of $\widehat{I}_n^{-1}$ (i.e. the **inverse** of the matrix $\widehat{I}_n$). The test statistic T_n under the null hypothesis converges to a standard normal distribution $\mathsf{N}(0,1)$.

The (asymptotic) confidence interval for θ_{Xk} is given by

$$\left(\widehat{\theta}_{nk} - \tfrac{u_{1-\alpha/2}\sqrt{i_n^{kk}}}{\sqrt{n}}, \widehat{\theta}_{nk} + \tfrac{u_{1-\alpha/2}\sqrt{i_n^{kk}}}{\sqrt{n}}\right).$$

Task 1

Let $(X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ be independent identically distributed random vectors. Suppose that the conditional density of Y_1 given X_1 is

$$f_{Y|X}(y|x;\beta) = \beta\, x\, \mathrm{e}^{-\beta\, x\, y}\, \mathbb{I}\{y > 0\},$$

where $\beta > 0$ is an unknown parameter. Further suppose that the distribution of X_1 does not depend on β.

(i) Find the maximum likelihood estimator of β.
(ii) Construct a test of the null hypothesis $H_0 : \beta = \beta_0$ against the alternative $H_0 : \beta \neq \beta_0$.
(iii) Construct a confidence interval for β.

Task 2

Suppose that you observe independent identically distributed random vectors $(X_1, Y_1)^{\mathsf{T}}, \ldots, (X_n, Y_n)^{\mathsf{T}}$ such that

$$\mathsf{P}(Y_1 = 1 \,|\, X_1) = \frac{\exp\{\alpha + \beta\, X_1\}}{1 + \exp\{\alpha + \beta\, X_1\}}, \qquad \mathsf{P}(Y_1 = 0 \,|\, X_1) = \frac{1}{1 + \exp\{\alpha + \beta\, X_1\}},$$

where the distribution of X_1 does not depend on the unknown parameters α a β.

(i) Derive a test for the null hypothesis $H_0 : \beta = 0$ against the alternative that $H_1 : \beta \neq 0$.
(ii) Find the confidence interval for the parameter β.