

MATH 523: SOME KEY ASYMPTOTIC RESULTS

In a probability model with independent data $y_1, \dots, y_n$, maximum likelihood estimation proceeds by solving the likelihood equations, obtained by setting the first derivative of the log-likelihood with respect to the parameters equal to zero. If $\theta = (\theta_1, \dots, \theta_p)^\top$ is the parameter vector, and

$$l_n(\theta) = l(\theta; \mathbf{y}) = \sum_{i=1}^n l(\theta; y_i) \quad (1)$$

is the log-likelihood, then

$$\dot{l}_n(\theta) = \dot{l}(\theta; \mathbf{y}) = \sum_{i=1}^n \dot{l}(\theta; y_i) = \sum_{i=1}^n \frac{\dot{f}_Y(x_i; \theta)}{f_Y(y_i; \theta)} = \sum_{i=1}^n S(y_i; \theta_0) \quad (2)$$

say, where $\dot{l}(\theta; y_i)$ is the $(p \times 1)$ vector of first partial derivatives, with j th element

$$\frac{\partial l(\theta; y_i)}{\partial \theta_j}$$

The function $\dot{l}(\theta; y_i) = S(y_i; \theta_0)$ is referred to as the **score function**. To compute the estimates, we must solve the $p \times 1$ system of **likelihood equations**

$$\dot{l}_n(\theta) = \dot{l}_n(\theta; \mathbf{y}) = 0_p \quad (\text{LE})$$

To understand the statistical properties of the resulting estimates and estimators, we consider the sampling distribution of the estimators assuming that the data are generated from the true distribution with parameter θ_0 , denoted $f_Y(y; \theta_0)$.

Recall that, under standard regularity conditions, if $Y_i \sim f_Y(y; \theta_0)$, then

$$\mathbb{E}_{f_Y}[\dot{l}(\theta_0; Y_i)] = \mathbb{E}_{f_Y}[S(\theta_0; Y_i)] = 0$$

Also, denote by $\mathcal{I}(\theta_0)$ the $p \times p$ matrix

$$\mathbb{E}_{f_Y}[S(\theta_0; Y_i)S(\theta_0; Y_i)^\top]$$

which can be shown to equal

$$-\mathbb{E}_{f_Y}[\Psi(\theta_0; Y_i)]$$

say, where $\Psi(\theta_0; y)$ is the matrix of second derivatives with (j, k) th element

$$\frac{\partial^2 l(\theta; y)}{\partial \theta_j \partial \theta_k} \quad j, k = 1, \dots, p.$$

This matrix is symmetric, and under regularity conditions, positive definite. The matrix $\mathcal{I}(\theta_0)$ is termed the **Fisher Information**.

Finally, recall the following distributional results: first, if $Y \sim \mathcal{N}(0, \sigma^2)$, then $(Y/\sigma)^2 \sim \chi_1^2$. Secondly, the multivariate equivalent of this result: if $Y \sim \mathcal{N}_p(0, \Sigma)$, then

$$Y^\top \Sigma^{-1} Y \sim \chi_p^2.$$

Consistency and Asymptotic Normality: Univariate Case

In the case $d = 1$, it is quite straightforward to show that a solution, $\hat{\theta}_n$, to the equation (LE) is *weakly consistent* for θ_0 , and *asymptotically normally distributed* provided that the log-likelihood is suitably differentiable with respect to θ on parameter space Θ , and otherwise regular.

- **Consistency:** As $n \rightarrow \infty$, we have that

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

- **Asymptotic Normality:** Consider a Taylor expansion of $\dot{l}_n(\theta)$ around θ_0

$$\dot{l}_n(\theta) = \dot{l}_n(\theta_0) + (\theta - \theta_0)\ddot{l}_n(\theta_0) + (\theta - \theta_0)^2 \ddot{l}_n(\theta^*)/2$$

where θ^* lies between θ_0 and θ . Evaluating this at $\theta = \hat{\theta}_n$, a root of the likelihood equation

$$0 = \dot{l}_n(\hat{\theta}_n) = \dot{l}_n(\theta_0) + (\hat{\theta}_n - \theta_0)\ddot{l}_n(\theta_0) + (\hat{\theta}_n - \theta_0)^2 \ddot{l}_n(\theta_n^*)/2$$

so that on rearrangement

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\dot{l}_n(\theta_0)/\sqrt{n}}{-(1/n)\ddot{l}_n(\theta_0) - (1/2n)(\hat{\theta}_n - \theta_0)\ddot{l}_n(\theta_n^*)}$$

Now, if

$$V(\theta_0) = \text{Var}_{f_Y}[S(Y; \theta_0)] = \mathcal{I}(\theta_0),$$

defines the **Fisher Information** $\mathcal{I}(\theta_0)$, then as $n \rightarrow \infty$, by the **Central Limit Theorem**,

$$\frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) = \sqrt{n} \frac{1}{n} \left\{ \sum_{i=1}^n \frac{\dot{f}_Y(Y_i|\theta_0)}{f_Y(Y_i|\theta_0)} \right\} \xrightarrow{d} Z \sim \mathcal{N}(0, V(\theta_0))$$

Similarly, by the Weak Law of Large Numbers, as $n \rightarrow \infty$,

$$-\frac{1}{n} \ddot{l}_n(\theta_0) = -\frac{1}{n} \sum_{i=1}^n \ddot{l}(\theta_0; Y_i) \xrightarrow{p} \mathcal{I}(\theta_0).$$

Finally,

$$\frac{1}{n}(\hat{\theta}_n - \theta_0)\ddot{l}_n(\theta_n^*) \xrightarrow{p} 0. \quad \therefore \quad \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

These results extend easily to the vector parameter case: if θ is a p -dimensional vector parameter, then the maximum likelihood estimate is still weakly consistent, and

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

Observed Information

For practical purposes, it is often necessary to produce an estimator of $\mathcal{I}(\theta_0)$. The *observed information*, $\hat{I}_n = \hat{I}_n(\mathbf{Y})$, can be used to approximate $\mathcal{I}(\theta_0)$, where any of the following definitions may be used.

$$\hat{I}_n = \begin{cases} \mathcal{I}(\hat{\theta}_n) \\ \frac{1}{n} \sum_{i=1}^n S(y_i; \hat{\theta}_n) S(y_i; \hat{\theta}_n)^\top \\ -\frac{1}{n} \sum_{i=1}^n \Psi(\hat{\theta}_n; y_i) \end{cases}$$

THE LIKELIHOOD RATIO TEST

Recall that the **Likelihood Ratio Test (LRT)** statistic for testing H_0 against H_1

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

is based on the statistic

$$\lambda(\mathbf{y}) = \frac{\sup_{\theta \in \Theta_0} f_{\mathbf{Y}}(\mathbf{y}|\theta)}{\sup_{\theta \in \Theta} f_{\mathbf{Y}}(\mathbf{y}|\theta)} = \frac{L(\hat{\theta}_{n0} | \mathbf{y})}{L(\hat{\theta}_n | \mathbf{y})}$$

say, where $\Theta \equiv \Theta_0 \cup \Theta_1$; H_0 is **rejected** if $\lambda(\mathbf{y})$ is **small enough**, that is, $\lambda(\mathbf{y}) \leq k$ for some k to be defined. **Note that Θ is not necessarily the entire parameter space, just the union of Θ_0 and Θ_1 .**

Asymptotic Behaviour of the Likelihood Ratio Test Statistic under the Null Hypothesis

Theorem Consider using the likelihood ratio statistic $\lambda(\mathbf{Y})$ for testing

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned} \tag{3}$$

Then under regularity assumptions, the maximum likelihood estimator for θ under H_1 , $\hat{\theta}_n$, is consistent and asymptotically normally distributed, and as $n \rightarrow \infty$ under H_0 ,

$$-2 \log \lambda(\mathbf{Y}) = -2(l_n(\theta_0) - l_n(\hat{\theta}_n)) \xrightarrow{d} Q \sim \chi_1^2$$

Proof A second-order Taylor expansion to the log-likelihood around the MLE, $\hat{\theta}_n$

$$l_n(\theta) = l_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n) \dot{l}_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) + \frac{1}{6}(\theta - \hat{\theta}_n)^3 \ddot{\ddot{l}}_n(\theta_n^*)$$

where θ_n^* lies between θ and $\hat{\theta}_n$. Now as $\hat{\theta}_n$ is the maximum likelihood estimate $\dot{l}_n(\hat{\theta}_n) = 0$, and therefore on rearrangement, evaluating at $\theta = \theta_0$,

$$-2(l_n(\theta_0) - l_n(\hat{\theta}_n)) = -(\theta_0 - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) - \frac{1}{3}(\theta_0 - \hat{\theta}_n)^3 \ddot{\ddot{l}}_n(\theta_n^*).$$

In this expression, consider the asymptotic behaviour in terms of the random variables. We have

$$-(\theta_0 - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) = (\sqrt{n}(\theta_0 - \hat{\theta}_n))^2 \left(-\frac{1}{n} \ddot{l}_n(\hat{\theta}_n) \right)$$

where

$$\sqrt{n}(\theta_0 - \hat{\theta}_n) \xrightarrow{d} Z \sim \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}) \quad -\frac{1}{n} \ddot{l}_n(\hat{\theta}_n) \xrightarrow{p} \mathcal{I}(\theta_0)$$

so that, by a standard theorem (**Slutsky's Theorem**)

$$-(\theta_0 - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) \xrightarrow{d} \chi_1^2$$

Finally,

$$(\theta_0 - \hat{\theta}_n)^3 \ddot{\ddot{l}}_n(\theta_n^*) = (\theta_0 - \hat{\theta}_n) \left((\theta_0 - \hat{\theta}_n)^2 \ddot{\ddot{l}}_n(\theta_n^*) \right) \xrightarrow{p} 0$$

so

$$-2(l_n(\theta_0) - l_n(\hat{\theta}_n)) \xrightarrow{d} Q \sim \chi_1^2$$

p -dimensional parameters

Consider using the likelihood ratio statistic for testing H_0 versus H_1 if θ is p -dimensional parameter. Under regularity conditions, as $n \rightarrow \infty$ under H_0 ,

$$-2 \log \lambda(\mathbf{Y}) = -2(l_n(\theta_0) - l_n(\hat{\theta}_n)) \xrightarrow{d} Q \sim \chi_p^2$$

This follows using identical methods to the $p = 1$ case. A second-order Taylor expansion to the log-likelihood around the MLE, $\hat{\theta}_n$, yields

$$l_n(\theta) \simeq l_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^\top \dot{l}_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^\top \ddot{l}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)$$

As $\hat{\theta}_n$ is the maximum likelihood estimate

$$\dot{l}_n(\hat{\theta}_n) = 0_p$$

and therefore on rearrangement, evaluating at $\theta = \theta_0$,

$$-2(l_n(\theta_0) - l_n(\hat{\theta}_n)) \simeq -(\theta - \hat{\theta}_n)^\top \ddot{l}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)$$

But, by previous results

$$(\theta - \hat{\theta}_n)^\top \ddot{l}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) = \sqrt{n}(\theta - \hat{\theta}_n)^\top \left(-\frac{1}{n} \ddot{l}_n(\hat{\theta}_n) \right) \sqrt{n}(\theta - \hat{\theta}_n) \xrightarrow{d} \mathbf{Z}^\top \mathcal{I}(\theta_0) \mathbf{Z}$$

as

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathbf{Z} \sim \mathcal{N}_p(0_p, \mathcal{I}(\theta_0)^{-1}) \quad -\frac{1}{n} \ddot{l}_n(\hat{\theta}_n) \xrightarrow{p} \mathcal{I}(\theta_0)$$

Hence, by standard distributional results, as $\mathbf{Z}^\top \mathcal{I}(\theta_0) \mathbf{Z} \stackrel{d}{=} \mathbf{X}^\top \mathbf{X}$ where $\mathbf{X} \sim \mathcal{N}_p(0_p, \mathbf{I}_p)$, we have

$$-2(l_n(\theta_0) - l_n(\hat{\theta}_n)) \xrightarrow{d} Q \sim \chi_p^2.$$

Composite Hypotheses

Theorem Consider testing the hypothesis

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

using the likelihood ratio statistic

$$\lambda(\mathbf{y}) = \frac{\sup_{\theta \in \Theta_0} f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)}{\sup_{\theta \in \Theta} f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)} = \frac{L(\hat{\theta}_{n0} | \mathbf{y})}{L(\hat{\theta}_n | \mathbf{y})}$$

say, where $\Theta \equiv \Theta_0 \cup \Theta_1$, and where H_0 specifies a model with k_1 free parameters (parameters not determined by the hypothesis), and H_1 and specifies a model with k_2 free parameters, with $k_2 > k_1$.

Then, under regularity conditions, as $n \rightarrow \infty$, under H_0

$$-2 \log \lambda(\mathbf{Y}) = -2(l_n(\hat{\theta}_{n0}) - l_n(\hat{\theta}_n)) \xrightarrow{d} Q \sim \chi_{k_2 - k_1}^2$$

Note: Such hypotheses can often be specified in the form

$$\begin{aligned} H_0 &: \theta = (\theta_0, \theta_1), \quad \theta_1 \text{ unspecified} \\ H_1 &: \theta \neq (\theta_0, \theta_1), \quad \theta_0, \theta_1 \text{ unspecified} \end{aligned}$$

that is, H_0 places constraints on one component θ , but leaves the other unspecified.