MATH 523: SOME KEY ASYMPTOTIC RESULTS

In a probability model with independent data $y_1, \ldots, y_n$, 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 $\theta = (\theta_1, \ldots, \theta_p)^\mathsf{T}$ is the parameter vector, and

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

is the log-likelihood, then

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

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

$$\dot{l}_n(\theta) = \dot{l}_n(\theta; \boldsymbol{y}) = 0_p \tag{LE}$$

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

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

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

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

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 $(Y/\sigma)^2 \sim \chi^2_1$. Secondly, the multivariate equivalent of this result: if $Y \sim \mathcal{N}_p(0, \Sigma)$, then

$$Y^\mathsf{T} \Sigma^{-1} Y \sim \chi^2_p.$$

Consistency and Asymptotic Normality: Univariate Case

In the case $d = 1$, it is quite straightforward to show that a solution, $\widehat{\theta}_n$, 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 $n \longrightarrow \infty$, we have that

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

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

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

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

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

so that on rearrangement

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

Now, if

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

defines the **Fisher Information** $\mathcal{I}(\theta_0)$, then as $n \longrightarrow \infty$, 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 $n \longrightarrow \infty$,

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

Finally, under the assumption that the third derivative $\dddot{l}_n(\theta)$ is bounded in expectation, we have

$$\frac{1}{n}(\widehat{\theta}_n - \theta_0)\dddot{l}_n(\theta_n^\star) \xrightarrow{p} 0 \qquad \therefore \qquad \sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

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

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

Observed Information

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

$$\widehat{I}_n = \begin{cases} \mathcal{I}(\widehat{\theta}_n) \\ \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} S(y_i;\widehat{\theta}_n)S(y_i;\widehat{\theta}_n)^{\mathsf{T}} \\ \\ -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\Psi(\widehat{\theta}_n; y_i) \end{cases}$$

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

$$\lambda(\boldsymbol{y}) = \frac{\sup\limits_{\theta \in \Theta_0} f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)}{\sup\limits_{\theta \in \Theta} f_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)} = \frac{L(\widehat{\theta}_{n0} \mid \boldsymbol{y})}{L(\widehat{\theta}_n \mid \boldsymbol{y})}$$

say, where $\Theta \equiv \Theta_0 \cup \Theta_1$; H_0 is **rejected** if $\lambda(\boldsymbol{y})$ is **small enough**, that is, $\lambda(\boldsymbol{y}) \leq k$ 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 $\lambda(\boldsymbol{Y})$ 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, $\widehat{\theta}_n$, is consistent and asymptotically normally distributed, and as $n \longrightarrow \infty$ under H_0,

$$-2 \log \lambda(\boldsymbol{Y}) = -2(l_n(\theta_0) - l_n(\widehat{\theta}_n)) \xrightarrow{d} Q \sim \chi^2_1$$

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

$$l_n(\theta) = l_n(\widehat{\theta}_n) + (\theta - \widehat{\theta}_n)\dot{l}_n(\widehat{\theta}_n) + \frac{1}{2}(\theta - \widehat{\theta}_n)^2\ddot{l}_n(\widehat{\theta}_n) + \frac{1}{6}(\theta - \widehat{\theta}_n)^3\dddot{l}_n(\theta^\star_n)$$

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

$$-2(l_n(\theta_0) - l_n(\widehat{\theta}_n)) = -(\theta_0 - \widehat{\theta}_n)^2\ddot{l}_n(\widehat{\theta}_n) - \frac{1}{3}(\theta_0 - \widehat{\theta}_n)^3\dddot{l}_n(\theta^\star_n).$$

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

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

where

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

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

$$-(\theta_0 - \widehat{\theta}_n)^2\ddot{l}_n(\widehat{\theta}_n) \xrightarrow{d} \chi^2_1$$

Finally,

$$(\theta_0 - \widehat{\theta}_n)^3\dddot{l}_n(\theta^\star_n) = (\theta_0 - \widehat{\theta}_n)\left((\theta_0 - \widehat{\theta}_n)^2\dddot{l}_n(\theta^\star_n)\right) \xrightarrow{p} 0$$

so

$$-2(l_n(\theta_0) - l_n(\widehat{\theta}_n)) \xrightarrow{d} Q \sim \chi^2_1$$

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 $n \longrightarrow \infty$ under H_0,

$$-2\log\lambda(\boldsymbol{Y}) = -2(l_n(\theta_0) - l_n(\widehat{\theta}_n)) \xrightarrow{d} Q \sim \chi^2_p$$

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

$$l_n(\theta) \doteq l_n(\widehat{\theta}_n) + (\theta - \widehat{\theta}_n)^{\mathsf{T}}\dot{l}_n(\widehat{\theta}_n) + \frac{1}{2}(\theta - \widehat{\theta}_n)^{\mathsf{T}}\ddot{l}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n)$$

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

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

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

$$-2(l_n(\theta_0) - l_n(\widehat{\theta}_n)) \doteq -(\theta - \widehat{\theta}_n)^{\mathsf{T}}\ddot{l}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n)$$

But, by previous results

$$(\theta - \widehat{\theta}_n)^{\mathsf{T}}\ddot{l}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n) = \sqrt{n}(\theta - \widehat{\theta}_n)^{\mathsf{T}}\left(-\frac{1}{n}\ddot{l}_n(\widehat{\theta}_n)\right)\sqrt{n}(\theta - \widehat{\theta}_n) \xrightarrow{d} \boldsymbol{Z}^{\mathsf{T}}\mathcal{I}(\theta_0)\boldsymbol{Z}$$

as

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

Hence, by standard distributional results, as $\boldsymbol{Z}^{\mathsf{T}}\mathcal{I}(\theta_0)\boldsymbol{Z} \overset{d}{=} \boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}$, where $\boldsymbol{U} \sim \mathcal{N}_p(0_p, \mathbf{I}_p)$, we have

$$-2(l_n(\theta_0) - l_n(\widehat{\theta}_n)) \xrightarrow{d} Q \sim \chi^2_p.$$

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

$$\lambda(\boldsymbol{y}) = \frac{\sup\limits_{\theta \in \Theta_0} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)}{\sup\limits_{\theta \in \Theta} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)} = \frac{L(\widehat{\theta}_{n0} \mid \boldsymbol{y})}{L(\widehat{\theta}_n \mid \boldsymbol{y})}$$

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 $n \longrightarrow \infty$, under H_0

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

Note: Such hypotheses can often be specified in the form

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

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