STAT-511 **2018 Spring**

Lecture 8: Estimation Equations

Instructor: Naveen N. Narisetty   Scribe: Yutong Dai   Last Modified: 2018-05-04

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications, and they may contain factual and/or typographic errors.*

8.1 Estimation Equations

Motivation When trying to find estimators of the parameter θ, we can use the maximum likelihood method, the moment method and etc. These methods, if work, can somehow be simplified to the equations. That is, to find the estimator, we indeed need to solve equations

$$G_n(\theta; \underset{\sim}{\mathbf{X}}) = 0,$$

where $G_n(.)$ can be a vector function and $\underset{\sim}{\mathbf{X}}$ represents the i.i.d samples. This motivates the definition of the estimation function and estimation equations.

Definition 8.1 Suppose $x_1, ..., x_n$ are i.i.d and $\boldsymbol{g_\theta}(x)$ is some continuous differentiable function of the data. Define the estimation function as

$$h_n(\theta) = \frac{1}{n}\sum_i \boldsymbol{g_\theta}(x_i),$$

and estimation equations as

$$h_n(\theta) = 0.$$

Definition 8.2 (unbiasedness) The estimation equation is said to be unbiased if

$$E_\theta[h_n(\theta)] = 0.$$

Remark This unbiasedness assumption is equivalent to assume $E_\theta(\boldsymbol{g_\theta}(x)) = 0$ for all θ.

The following subsection use the approach of estimation equations to address the **Misspecified Model** issue when doing the Maximum Likelihood.

8.1.1 Application: Misspecified Maximum Likelihood

Suppose the data comes from the model $\boldsymbol{f_\theta}$, that is $x_1, ..., x_n \overset{iid}{\sim} \boldsymbol{f_\theta}$. But somehow, we wrongly specify the $\boldsymbol{g_\theta}$ as the working likelihood. So we actually estimate the true θ by

$$\arg\max_\theta \sum_i \log \boldsymbol{g_\theta}(x_i) \tag{8.1}$$

Denote $\hat{\theta}$ as a solution of Equation 8.1, and

$$h_n(\theta) = \sum_i \frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x_i) \triangleq \ell'_n(\theta).$$

Also here, we assume $h_n(\theta) = 0$ is **unbiased**, which is equivalent to $E_\theta(\frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x)) = 0$

Consider the following Taylor expansion,

$$\ell'_n(\theta) = \ell'_n(\hat\theta) + (\theta - \hat\theta)\ell''_n(\tilde\theta),$$

which implies

$$\sqrt{n}(\hat\theta - \theta) = \sqrt{n}\frac{\ell'_n(\theta)}{\ell''_n(\tilde\theta)} \approx \sqrt{n}\frac{\ell'_n(\theta)}{\ell''_n(\theta)}$$

By CLT,

$$\sqrt{n}(\frac{1}{n}\ell'_n(\theta) - E(\frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x))) \xrightarrow{d} N(0, Var(\frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x))).$$

Combining with the unbiasedness assumption, we derive

$$\frac{1}{\sqrt{n}}\ell'_n(\theta) \xrightarrow{d} N(0, D_1),$$

where $D_1(\theta) = Var(\frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x))$.

By LLN,

$$\frac{1}{n}\ell''_n(\theta) = \frac{1}{n}\sum_i \frac{\partial^2}{\partial\theta^2} \log \boldsymbol{g_\theta}(x_i) \xrightarrow{p} E(\frac{\partial^2}{\partial\theta^2} \log \boldsymbol{g_\theta}(x_i)) \triangleq E(\ell''_1(\theta)).$$

Denote $-E(\ell''_1(\theta) \triangleq D_2(\theta)$.

So with the working likelihood, we can still establish the astmptotic normality for the estimator $\hat\theta$,

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N(0, D_2^{-1}(\theta)D_1(\theta)D_2^{-1}(\theta)).$$

Remark If $\boldsymbol{f_\theta}(x) = \boldsymbol{g_\theta}(x)$, then we have $D_1(\theta) = D_2(\theta)$. This is because

$$E[\frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x)] = \int_x \frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x)\boldsymbol{f_\theta}(x)dx = \int \boldsymbol{g'_\theta}(x)dx = \frac{\partial}{\partial\theta}\int \boldsymbol{g_\theta}(x)dx = 0 \tag{8.2}$$

Set derivative with respect to θ on both sides of Equation 8.2, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int_x \frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x)\boldsymbol{f_\theta}(x)dx \\ &= \int_x \frac{\partial^2}{\partial\theta^2} \log \boldsymbol{g_\theta}(x)\boldsymbol{f_\theta}(x)dx + \int_x \frac{\partial}{\partial\theta} \log \boldsymbol{g_\theta}(x)\boldsymbol{f'_\theta}(x)dx \\ &= \int_x (\ell''_n(\theta) + \ell'_n(\theta)\frac{\boldsymbol{f'_\theta}(x)}{\boldsymbol{f_\theta}(x)})\boldsymbol{f_\theta}(x)dx \end{aligned} \tag{8.3}$$

This implies $D_1(\theta) = D_2(\theta)$. Then we are back to the normality property under the likelihood framework.

RemarK The $D_2^{-1}(\theta)D_1(\theta)D_2^{-1}(\theta)$ is called **Godambe Information**. When choosing the proper working likelihood, we prefer $\boldsymbol{g_\theta}(x)$ that gives larger Godambe Information, hence small variance.

8.1.2 Genearlization: Asymptotic Distribution

Suppose $\hat{\theta}$ is the root of $h_n(\theta) = \frac{1}{n}\sum_i \boldsymbol{g_\theta}(x_i) = 0$ and the estimation equation is unbiased, then we have

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, D_2^{-1}(\theta)D_1(\theta)D_2^{-1}(\theta)),$$

where $D_1(\theta) = Var(\boldsymbol{g_\theta}(x_1))$ and $D_2 = -E(\frac{\partial}{\partial\theta}\boldsymbol{g_\theta}(x_i))$.

8.2 Weak Law for Random Functions

Suppose $x_1, ..., x_n \overset{iid}{\sim} \boldsymbol{f_\theta}$, where $\theta \in K$ and K is a compact set in R^p. Consider a sequence of functions $W(\theta, x_i)$, which are i.i.d. random functions taking value in the function space $C(K)$, the space of function continuous functions on K.

Remark: Here, we regard $W(\theta, x_i)$ as a function of the θ, and the randomness comes from introducing the data x_i.

A natural question is whether $\frac{1}{n}\sum_i W(\theta, x_i) \xrightarrow{p} E(W(\theta, x_1))$?

By LLN, for each fixed θ, $\frac{1}{n}\sum_i W(\theta, x_i) \xrightarrow{p} E(W(\theta, x_1))$, which is point-wise convergence. The following theorem establishes the uniform convergence.

Definition 8.3 (infinity norm) For any $W(\theta, x) \in C(K)$, $||W(\theta, x)||_\infty = \sup_{\theta\in K} W(\theta, x)$.

Theorem 8.4 (uniform convergence) Suppose $W(\theta, x_1), ..., W(\theta, x_n)$ be i.i.d. random funcitons with $E||W(\theta, x_i)||_\infty < +\infty, \forall i$. Then

$$||\frac{1}{n}\sum_i W(\theta, x_i) - E(W(\theta, x_1))||_\infty \xrightarrow{p} 0$$

The following theorem shows how Theorem 8.4 can be useful.

Theorem 8.5 If for any random function $G_n(\theta, \underset{\sim}{\mathbf{X}})$ in $C(K)$, we have

$$||G_n(\theta, \underset{\sim}{\mathbf{X}}) - G(\theta)||_\infty \xrightarrow{p} 0,$$

for a nonrandom function $G(\theta)$ in $C(K)$. Then

1. If $\hat{\theta}_n \xrightarrow{p} \theta^*$, then $G_n(\hat{\theta}_n, \underset{\sim}{\mathbf{X}}) \xrightarrow{p} G(\theta^*)$.

2. If θ^* is the unique maximizer of $G(\theta)$ and $\hat{\theta}_n \in \arg\max_\theta G_n(\theta, \underset{\sim}{\mathbf{X}})$, then $\hat{\theta}_n \xrightarrow{p} \theta^*$.

3. If θ^* is the unique root of $G(\theta) = 0$ and $\hat{\theta}_n$ is a root of $G_n(\theta, \underset{\sim}{\mathbf{X}}) = 0$, then $\hat{\theta}_n \xrightarrow{p} \theta^*$.