Chapter 11 Advanced Estimation Theory

► Goldberger, Ch. 12

Yale Note Ch. 15, 16

Wackerly et al. Ch. 9-7, 9-8

1. Method of Moment Estimation

• Consider a population in which random variable X has pdf $f(x;\theta)$ with the function $f(\cdot)$ known and the parameter θ unknown.

► Earlier, we defined rth moment of a random variable X, $\mu_r' = E(X^r)$.

► If $X \sim f(x;\theta)$, actually, $\mu_r' = \mu_r'(\theta)$.

• Idea of Method-of-Moment Estimator: θ can be estimated by equating the true moments μ_r' and the corresponding sample moments

$$\hat{\mu}_r'(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i^r(\theta)$$

and solving the resulting equations for the unknown parameters θ.

(Example) Let X_i be a random sample from $N(\mu, \sigma^2)$ population, $\theta = (\mu, \sigma^2)$.

Since $\quad (\mu_1' =)\ E(X) = \mu$,

$\qquad (\mu_2' =)\ E(X^2) = \mu^2 + \sigma^2$,

and $\quad \hat{\mu}_1' = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$\hat{\mu}_2' = \frac{1}{n}\sum_{i=1}^{n} X_i^2 .$$

Therefore,

$$\hat{\mu}_1' = \hat{\mu} \qquad \Rightarrow \qquad \hat{\mu} = \bar{X}$$

$$\hat{\mu}_2' = \hat{\sigma}^2 + \hat{\mu}^2 \qquad \Rightarrow \qquad \hat{\sigma}^2 = \hat{\mu}_2' - \hat{\mu}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

- (Definition) Given a function $m(.)$ such that $E[m(X;\theta)]=0$ iff $\theta=\theta_0$, then Method-of-Moment estimator $\hat{\theta}$ solves $\frac{1}{n}\sum_{i=1}^{n} m(X_i;\hat{\theta})=0$.

- This methodology applies in principle in the case that there are r parameters involved $\theta_1, \theta_2, \cdots, \theta_r$, $r \geq 1$. In this case we have to estimate that the r first moments of the X_i's; that is,

$$E(X_i^k)=m_k(\theta_1,\cdots,\theta_r), \qquad k=1,2,\cdots,r.$$

Then form the first r sample moments;

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = m_k(\hat{\theta}_1,\cdots,\hat{\theta}_r), \qquad k=1,2,\cdots,r.$$

◎ <u>Example of moment equations</u>

① X has mean μ.

$$E(X) = \mu \quad \Rightarrow \quad E(X - \mu) = 0$$

② X has variance σ^2.

$$E(X - \mu)^2 = \sigma^2 \quad \Rightarrow \quad E\left[(X - \mu)^2 - \sigma^2 \right] = 0$$

③ $Y_t \sim AR(1)$: $Y_t = \theta_0 + \theta_1 Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim$ white noise.

Then,

$$E(Y_t) = \frac{\theta_0}{1-\theta_1}$$

$$V(Y_t) = \frac{\sigma^2}{(1-\theta_1)^2}$$

$$Cov(Y_t, Y_{t-1}) = \theta_1 V(Y_t)$$

④ Linear Least Squares Estimator: $y = X\beta + \varepsilon$.

$E(X'\varepsilon) = 0$ or $E\left[X'(y - X\beta)\right] = 0$.

(Example) Intertemporal Asset Pricing Model

Individual maximizes $E\left[\sum_{s=0}^{S}\delta^{s}U(\mathrm{c}_{t+s})\right]$

subject to $c_{t+s}+q_{t+s}=w_{t+s}+(1+r_{t+s})q_{t+s-1}$.

► Here, $U=U(\theta)$, we are estimating θ.

F.O.C.:

$$E\left[\delta U'(c_{t+1})(1+r_{t+1})\right]=U'(c_t)$$
$$\Rightarrow E\left[\delta\frac{U'(c_{t+1})}{U'(c_t)}(1+r_{t+1})-1\right]=0$$
.

► Replace the population moment with sample moment and solve for $\hat{\theta}$.

2. Maximum Likelihood Estimation

• Consider a population in which random variable X has pdf $f(x;\theta)$ with the function $f(\cdot)$ known and the parameter θ unknown.

(1) Example: Discrete sample

• Consider tossing a crooked coin.

$X_i = 1 \ if \ head$ at ith tossing.

Then $X_i \sim Bernoulli(p)$ with $p = P(X_i = 1)$ is unknown.

► We want to estimate $p = P(X_i = 1)$.

► Suppose we toss it ten times and a head appears nine times

⇒ event A=(9H, 1T).

Since we have (9H, 1T) rather than (5H, 5T), $p = \frac{1}{2}$ is not likely.

►

$$P(A \mid p = 1/2) = C_9^{10} (\tfrac{1}{2})^{10} = 0.01$$
$$P(A \mid p = 3/4) = C_9^{10} (\tfrac{3}{4})^{9} \left(\tfrac{1}{4}\right) = 0.19$$
$$P(A \mid p = 9/10) = C_9^{10} (\tfrac{9}{10})^{9} \left(\tfrac{1}{10}\right) = 0.39$$

From this, we can conjecture that $p = \frac{9}{10}$ is more likely than $p = \frac{3}{4}$ or $p = \frac{1}{2}$.

► $P(A \mid p) = C_9^{10} p^9 (1-p)$: *likelihood function* of $p = P(X_i = 1)$.

- Under random sampling $\left(x_1,x_2,\cdots,x_n\right)$, the joint pdf for the sample is

$$P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)=\prod_{i=1}^{n}f(x_i;\theta).$$

► *Likelihood function*: $L\left(x_1,x_2,\cdots,x_n;\theta\right)=\prod_{i=1}^{n}f(x_i;\theta)$

► Our object is to estimate θ.

(Definition) *Maximum likelihood estimator* of θ is the value for θ that maximizes the likelihood (joint probability) function for θ, that is,

$$\hat{\theta}=\underset{\theta}{\operatorname{argmax}}\,L(x_1,x_2,\cdots,x_n;\theta).$$

► MLE means choosing that probability distribution (p in the above example) under which the observed values could have occurred with the highest probability.

(2) Continuous sample

- Modify the discrete case slightly.

(Definition) Let $(x_1, x_2, \cdots, x_n)$ be a random sample on a continuous population with a density function $f(x;\theta)$.

Then we call $L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$ the *likelihood function* of θ.

► The value of θ that maximizes L, the *maximum likelihood estimator*.

- MLE $\hat{\theta}$ maximizes $L(x_1, x_2, \cdots, x_n; \theta)$.

  Also, MLE $\hat{\theta}$ maximizes $\log L(x_1, x_2, \cdots, x_n; \theta)$.

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} L\left(x_1, x_2, \cdots, x_n; \theta\right) \\ &= \arg\max_{\theta} \log L\left(x_1, x_2, \cdots, x_n; \theta\right) \\ &= \arg\max_{\theta} \sum_{\mathrm{i}=1}^{\mathrm{n}} \log \mathrm{f}(x_i; \theta)\end{aligned}$$

► If $L(\theta_1) \geq L(\theta_2)$ for all θ_2, $\log L(\theta_1) \geq \log L(\theta_2)$.

Therefore, $\hat{\theta} = \arg\max_{\theta} \ L\left(x_1, x_2, \cdots, x_n; \theta\right) = \arg\max_{\theta} \ \log L\left(x_1, x_2, \cdots, x_n; \theta\right)$.

► $\log L(\mathbf{x}; \theta) = \log\left(\prod_{i=1}^{n} f(x_i; \theta)\right) = \sum_{i=1}^{n} \log f(x_i; \theta) \equiv \mathcal{L}_n(\mathbf{x}; \theta)$; <u>*log-likelihood function*</u>

F.O.C.: $\dfrac{\partial \mathcal{L}_n(\mathbf{x}; \hat{\theta})}{\partial \theta} = 0 = \sum_{i=1}^{n} \dfrac{\partial \log f(x_i; \theta)}{\partial \theta}$.

► Check S.O.C

(Example) Let $\left(y_1, y_2, \cdots, y_n\right)$ be random sample from Bernoulli distribution with $P(Y_i = 1) = p$.

☞ $\hat{p} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$.

(Example) Let $\left(y_1, y_2, \cdots, y_n\right)$ be random sample from $N\left(\mu, \sigma^2\right)$ $(\theta = (\mu, \sigma^2))$.

☞ $\hat{\mu} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$, $\hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2$.

(Note)

① $E(\hat{\mu}) = E(\overline{y}) = \mu$.

② $E(\sigma^2) = \dfrac{n-1}{n} E\left(\dfrac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 \right) = \dfrac{n-1}{n} \sigma^2 \neq \sigma^2$ but as $n \to \infty,\ \hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

(3) Computation

► The likelihood equation is often so highly nonlinear in the parameters, however, that it can be solved only by some method of iteration.

• Newton-Rapson method:

Quadratic Taylor expansion of $\mathcal{L}_n(\mathbf{x};\theta)$ around θ^*:

$$\left.\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\right|_{\hat{\theta}} = 0 = \left.\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\right|_{\theta^*} + \left.\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\right|_{\theta^*} (\theta - \theta^*).$$

So,

$$\theta = \theta^* - \left(\left.\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\right|_{\theta^*} \Bigg/ \left.\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\right|_{\theta^*} \right).$$

► Therefore, from initial value $\hat{\theta}_1$,

the second-round estimator $\hat{\theta}_2$ can be obtained as

$$\hat{\theta}_2 = \hat{\theta}_1 - \left(\left.\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\right|_{\hat{\theta}_1} \Bigg/ \left.\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\right|_{\hat{\theta}_1} \right).$$

The iteration should be repeated until it converges as follows:

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \left(\left.\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\right|_{\hat{\theta}_i} \Bigg/ \left.\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\right|_{\hat{\theta}_i} \right).$$

(4) Properties of MLE

① Consistency: $p\lim \hat{\theta} = \theta_0$

② Normality: $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \dfrac{1}{I(\theta_0)}\right)$.

Or,

$$\hat{\theta} \overset{a}{\sim} N\left(\theta_0, \frac{1}{n \cdot I(\theta_0)}\right) ,$$

where $I(\theta_0) = -E\left(\dfrac{d^2 \log f(y_i;\theta_0)}{d\theta^2}\right) = E\left[\left(\dfrac{d \log f(y_i;\theta_0)}{d\theta}\right)^2\right]$: Fisher Information

- $$A.V(\hat{\theta}) = -\frac{1}{n \cdot E\left(\frac{d^2 \log f(y_i;\theta_0)}{d\theta^2}\right)}$$

  ► Estimator of $E\left(\frac{d^2 \log f(y_i;\theta_0)}{d\theta^2}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{d^2 \log f(y_i;\hat{\theta})}{d\theta^2}$

- Estimator of $A.V(\hat{\theta}) = -\frac{1}{\sum_{i=1}^{n}\frac{d^2 \log f(y_i;\hat{\theta})}{d\theta^2}}$.

► Asymptotic Standard Error of $\hat{\theta} = \sqrt{-\frac{1}{\sum_{i=1}^{n}\frac{d^2 \log f(y_i;\hat{\theta})}{d\theta^2}}}$.

③ Cramer-Rao Inequality

Let $\tilde{\theta}=\tilde{\theta}(X_1,\cdots,X_n)$ be an consistent(or unbiased) estimator of θ. Then under general conditions, we have $A.V.(\tilde{\theta}) \geq -\dfrac{1}{n \cdot I(\theta_0)} = A.V.(\hat{\theta})$.

The right-hand side is known as the *Cramer-Rao lower bound(CRLB).*

⇒ MLE $\hat{\theta}$ is most efficient since no lower variance is possible for an unbiased (consistent) estimator.

④ Proof of Asymptotic Normality

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)=\left[-\frac{1}{\dfrac{d^2\mathcal{L}_n(\theta^*)}{d\theta^2}}-\right]\sqrt{n}\,\frac{d\mathcal{L}_n(\theta_0)}{d\theta}$$

$$=\left[-\frac{1}{\dfrac{d^2\log L(\theta^*)}{d\theta^2}}\right]\sqrt{n}\,\frac{d\log L(\theta_0)}{d\theta}$$

$$=\left[-\frac{1}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\dfrac{d^2\log f_i(\theta^*)}{d\theta^2}}\right]\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\frac{d\log f_i(\theta_0)}{d\theta} \qquad (*)$$

ⓐ (score function) For score function,

$$E\left(\frac{d\log f_i(\theta_0)}{d\theta}\right)=0.$$

(proof)

ⓑ (Hessian)

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{d^2\log f_i(\theta^*)}{d\theta^2}\xrightarrow{\quad p\quad}-E\left(\frac{d^2\log f_i(\theta_0)}{d\theta^2}\right)=I(\theta_0).$$

ⓒ (Fisher information) $I(\theta_0) \equiv -E\left(\frac{d^2 \log f_i(\theta_0)}{d\theta^2}\right) = E\left[\left(\frac{d \log f_i(\theta_0)}{d\theta}\right)^2\right]$.

(proof)

⑤ Proof of Efficiency

Consider other unbiased estimator $\tilde{\theta}$ such that $E(\tilde{\theta}) = \theta_0$.

From this property, we can get $E\left(\tilde{\theta} \cdot \sum_{i=1}^{n} \frac{d \log f(y_i;\theta)}{d\theta}\right) = 1$.

Since $E\left(\frac{d \log f(y_i;\theta)}{d\theta}\right) = 0 \quad \Rightarrow \quad E\left(\sum_{i=1}^{n} \frac{d \log f(y_i;\theta)}{d\theta}\right) = 0$,

$$Cov\left(\tilde{\theta}, \sum_{i=1}^{n} \frac{d \log f(y_i;\theta)}{d\theta}\right) = 1.$$

Apply Cauchy-Schwartz inequality, $V(X) - \frac{C(X,Y)^2}{V(Y)} \geq 0$, then, $V(\tilde{\theta}) \geq \dfrac{1}{V\left(\sum_i \dfrac{d \log f(y_i;\theta)}{d\theta}\right)} = \dfrac{1}{n \cdot E\left(\dfrac{d^2 \log f(y_i;\theta)}{d\theta^2}\right)} = V(\hat{\theta})$.

(Example) Let $\left(y_1, y_2, \cdots, y_n\right)$ be random sample from Bernoulli distribution with $P(Y_i = 1) = p$.

Find CRLB.