12. Model Construction

12.1 –Properties of Parameter Estimators

Unbiasedness

$E\left[\hat{\theta}\middle|\theta\right]=\theta,\forall\theta$

$Bias := E\left[\hat{\theta}\middle|\theta\right]-\theta$

Asymptotic Unbiasedness

$\lim_{n\to\infty} E\left[\hat{\theta}_n\middle|\theta\right]=\theta$

For a sample size *n*.

Uniformly Minimum Variance Unbiased Estimator (UMVUE)

Unbiasedness must hold for any sample size. An estimator is UMVUE if:

- It is unbiased
- For any true value of θ there is no other unbiased estimator with a smaller variance

(Weak) Consistency

$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n-\theta\right|>\delta\right)=0$

Sufficient conditions:

- Estimator is asymptotically unbiased
- $Var\left(\hat{\theta}_n\right)\to 0$

Unbiasedness is a property that refers to samples of all sizes; **consistency** is only applicable large samples only.

Uniformly Most Powerful (Hypothesis Test)

A hypothesis test is **uniformly most powerful if**:

- for a given significance level,
- no other test exists that can give a smaller probability of committing a Type 2 error (falsely not rejecting H_0)

MSE& Bootstrapped MSE

$MSE=E\left[\left(\hat{\theta}-\theta\right)^2\right]$

$=Var\left(\hat{\theta}\right)+bias\left(\hat{\theta}\right)^2$

$bias\left(\hat{\theta}\right)=E\left(\hat{\theta}\right)-\theta$

N.B.: $Var\left(\hat{\theta}\right)$ refers to *sample variance*, i.e. use 1/n, not 1/(n-1)

Bootstrap MSE

1. For a sample of size *n*, generate N = n^n empirical observations with replacement, to get x_i data points, *i* = 1…N.
2. Mass produce *N* possible values of the estimators θ
3. $MSE_{Bootstrap}=\frac{1}{N}Var\left(\hat{\theta}\right)=\sum_{i=1}^{N}p\left(\theta_i\right)\left(\theta_i-\bar{\theta}\right)$

12.2 – Interval Estimators

A 100(1 - α)% CI for some (unknown) true parameter θ is a pair of r.v. *L* and *U* (the interval estimator) s.t. at least 100(1 - α)% of the time over a variety of samples, [L, U] will enclose the true value.

1. CI for Student-t Distribution

Sample mean of *n* i.i.d. normally distributed r.v. X_i with same mean and variance ~t_{n-1}:

$X_i \sim N\left(\mu,!^{\,2}\right),i=1,2\ldots n$

$\bar{X}\in\mu\pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right),s=\sqrt{\frac{\sum_i^n\left(X_i-\mu\right)^2}{n-1}}$

$t_{\alpha/2,n-1}$ is the 100(1 - α/2)th percentile of the *t* distribution with *n-1* d.f.

2. Normally Distributed Estimators

If we know the 1st 2 moments of the parameter estimator, and that it is normally distributed, then:

$\hat{\theta}\sim N\left(\theta,Var\left(\theta\right)=\hat{v}\left(\theta\right)\right)$

$P\left(-z_{\alpha/2}\le\frac{\hat{\theta}-\theta}{\sqrt{\hat{v}\left(\theta\right)}}\le z_{\alpha/2}\right)=1-\alpha$

v(θ) may be difficult to find due to the presence of *θ* in the denominator.
For nice distributions, use CLT.

Converting Discrete to Continuous Data

Empirical CDF

$\hat{F}\left(x\right)=\frac{\#\,observations\le x}{n}$

Ogive (Grouped Data)

$\tilde{F}\left(x\right)=\frac{c_i-x}{c_i-c_{i-1}}\tilde{F}\left(c_{i-1}\right)+\frac{x-c_{i-1}}{c_i-c_{i-1}}\tilde{F}\left(c_i\right),c_{i-1}\le x\le c_i$

Kernel Density Estimators

Assigns the probability mass to a neighbourhood around x_i, rather than assigning it completely to a point. For a given interval [$x_i - b, x_i + b$], $b > 0$ is the **bandwidth**.

$\tilde{f}\left(y\right)=\sum_{x_i}f_n\left(x_i\right)\times f_{KDE}\left(x_i\right),x_i\in\left[y-b,y+b\right]$

$\tilde{F}\left(y\right)=\sum_{x_i}f_n\left(x_i\right)\times F_{KDE}\left(x_i\right),x_i\le\left(y+b\right)$

1st Moment of the smoothed pdf is the empirical 1st moment:

$E\left[X\right]=\frac{\sum x}{n}$

Uniform

$$K_U\left(y\right)=\begin{cases}0 & y<x_i-b\\ \frac{y-x_i+b}{2b} & x_i-b\le y\le x_i+b\\ 1 & y>x_i+b\end{cases},k_U\left(y\right)=\begin{cases}\frac{1}{2b} & x_i-b\le y\le x_i+b\\ 0 & otherwise\end{cases}$$

$Var\left(\tilde{X}\right)=Var\left(\ddot{X}_i\right)+\frac{b^2}{3}$

Triangular

$$K_\Delta\left(y\right)=\begin{cases}0 & y<x_i-b\\ \frac{\left(y-\left(x_i-b\right)\right)^2}{2b^2} & x_i-b\le y\le x_i\\ 1-\frac{\left(y-\left(x_i+b\right)\right)^2}{2b^2} & x_i\le y\le x_i+b\\ 1 & y>x_i+b\end{cases},k_\Delta\left(y\right)=\begin{cases}0 & y<x_i-b\\ \frac{y-x_i+b}{b^2} & x_i-b\le y\le x_i\\ 1-\frac{y-x_i+b}{b^2} & x_i\le y\le x_i+b\\ 1 & y>x_i+b\end{cases}$$

$Var\left(\tilde{X}\right)=Var\left(\hat{X}_i\right)+\frac{b^2}{6}$

14. Estimation with Incomplete Data

Risk Sets

$$r_j = \#\left(d_i : d_i < y_j\right) - \#\left(u_i : u_i < y_j\right) - \#\left(x_i : x_i < y_j\right)$$

$$r_j = r_{j-1} - s_{j-1} + \#\left(d_i : y_{j-1} \le d_i < y_j\right) - \#\left(u_i : y_{j-1} \le u_i < y_j\right)$$

- $\{x_i\}_{i = 1\ldots n}$:= *n* uncensored observations. *i* is the index of each individual observed datum, where *i = 1…n*.
- $\{y_j\}_{j = 1\ldots k}$:= *k* unique values of observed values. *j* is the index of each unique observed datum, where *j = 1…k*.
- u_i:= (right) censored observations (withdrawals)
- d_i:= (left) truncated observations (entries midway into the study)
- s_j := no. of times the observation y_j appears.
- r_j := sample size of risk set *j* (the set that comprises the individuals under observation at the time of study)

Kaplan-Meier/Product-Limit

$$\hat{S}_{KM}(y) = \prod_{j=1}^{i}\left(1 - \frac{s_j}{r_j}\right)$$

$$y \in \left[y_j, y_{j+1}\right) j = 2\ldots m$$

Moments

$$P\left(X \le y_i \middle| X > y_{i-1}\right) = \frac{S(y_{i-1}) - S(y_i)}{S(y_{i-1})} = 1 - S_i \equiv \frac{\int_{y_{i-1}}^{y_i} f(y)\,dy}{1 - F(y_{i-1})}$$

$$E\left[\hat{S}_{KM}(y)\right] = \prod_{i=1}^{j}\frac{\hat{S}_{KM}(y_i)}{\hat{S}_{KM}(y_{i-1})}$$

$$Var\left[\hat{S}_{KM}(y)\right] = S(y_i)^2\left\{\prod_{i=1}^{j}\left(\frac{1-S_i}{S_i r_i} + 1\right) - 1\right\} \approx S(y_i)^2\sum_{i=1}^{j}\frac{1-S_i}{S_i r_i} = \hat{S}_{KM}(y_i)^2\sum_{i=1}^{j}\frac{s_i}{r_i(r_i - s_i)}$$

The approximation is known as *Greenwood's approximation.*

Nelson-Aalen

$$\hat{H}_{NA}(y) = \begin{cases} 0 & y < y_1 \\ \sum_{j=1}^{m}\frac{s_i}{r_i} & y \in \left[y_i, y_{i+1}\right) \\ \sum_{j=1}^{i}\frac{s_i}{r_i} & y > y_m \end{cases} \xleftrightarrow{S(x)=e^{-H(x)}} \hat{S}_{NA}(y) = \begin{cases} 0 & y < y_1 \\ \exp\left[-\sum_{j=1}^{m}\frac{s_i}{r_i}\right] & y \in \left[y_i, y_{i+1}\right) \\ \exp\left[-\sum_{j=1}^{i}\frac{s_i}{r_i}\right] & y > y_m \end{cases}$$

Moments

$$Var\left[\hat{H}(y)\right] = \sum_{i=1}^{j}\frac{s_i}{r_j^2} \xrightarrow{Delta-Method} \left(\frac{d\hat{S}(y)}{d\hat{H}(y)}\right)^2 Var\left[\hat{H}(y)\right] = e^{-2\hat{H}(y)}Var\left[\hat{H}(y)\right]$$

$$\hat{S}(y) = e^{-\hat{H}(y)}, \frac{d\hat{S}(y)}{d\hat{H}(y)} = -e^{-\hat{H}(y)}$$

S Confidence Intervals

Linear	95% C.I. of y_i	$\hat{S}_{KM}(y_i) \pm 1.96\sqrt{Var\left(\hat{S}_{KM}(y_i)\right)}$
Log*	95% C.I. of δ_i δ_i = *lg[-lg(y)]*	$\left(\hat{S}(y)^U, \hat{S}(y)^{1/U}\right)$ $U = \exp\left[1.96\frac{\sqrt{Var\left(\hat{S}(y)\right)}}{\left[\hat{S}(y)\lg\left(\hat{S}(y)\right)\right]}\right]$

*This is the **Delta-Method**: for an estimator

$$If : \hat{\theta} \sim N\left(\theta, \sigma^2\right)$$

$$Then : g\left(\hat{\theta}\right) \sim N\left(g(\theta), \left(g'(\theta)^2\sigma^2\right)\right)$$

Problem Type

$$Given : \left(\hat{S}(y)^U, \hat{S}(y)^{1/U}\right)$$

1. $U\lg\left(\hat{S}\right) \times \frac{1}{U}\lg\left(\hat{S}\right) = \left[\lg\left(\hat{S}\right)\right]^2, U^2 = \frac{\frac{1}{U}\lg\left(\hat{S}\right)}{U\lg\left(\hat{S}\right)}$
2. $\lg\left(\hat{S}\right) = \ a \quad OR \quad -a$
3. $Choose : \hat{S} = e^{-a}, s.t. \hat{S} \in [0,1]$

H Confidence Intervals

95% C.I. of *y* *y* ∈ *[yi, yi+1)*	$\hat{H}(y_i) \pm 1.96\sqrt{Var\left(\hat{H}(y_i)\right)}$
(1 - α)% C.I. of δ_i δ_i = *lg[-lg(y)]*	$\left(\hat{H}(y)U, \hat{H}(y)/U\right)$ $U = \exp\left[z_{1-\alpha/2}\frac{\sqrt{Var\left(\hat{H}(y_i)\right)}}{\hat{H}(y_i)}\right]$

(Kaplan-Meier) Approximation for Large Data Sets

Assume that all the truncation *(d)* and censoring *(u)* takes place uniformly throughout the interval. Then:

$$r_0 = \frac{(d_0 - u_0)}{2}$$

$$r_j = \frac{(d_j - u_j)}{2} + \sum_{i=0}^{j-1}(d_i - u_i - x_i), j = 1,2\ldots k$$

Interval Estimation (Normal Approximation)

$$F_n(y) = \frac{Y}{n} \to \#obs \le Y \to Y \sim Bin\left(m = n, q = \tilde{F}(y)\right)$$

$$E\left[\tilde{F}(y)\right] = \tilde{F}(y), Var\left(\tilde{F}(y)\right) = \frac{\tilde{F}(y)\left(1 - \tilde{F}(y)\right)}{n}$$

$$Interval:$$

$$\tilde{F}(y) \pm z_{1-\alpha/2}\sqrt{\frac{\tilde{F}(y)\left(1 - \tilde{F}(y)\right)}{n}}$$