Lecture 23: Variance estimation, replication, jackknife, and bootstrap

Motivation

To evaluate and compare different estimators, we need consistent estimators of variances or asymptotic variances of estimators.
This is also important for hypothesis testing and confidence sets.
Let Var($\widehat{\theta}$) be the variance or asymptotic variance of an estimator $\widehat{\theta}$.
Traditional approach to estimate Var($\widehat{\theta}$): Derivation and substitution

- First, we derive a theoretical formula
- Approximation (asymptotic theory) is usually needed
- The formula may depend on unknown quantities
- We then substitute unknown quantities by estimators

Example: the δ-method

$Y_1, ..., Y_n$ are iid (k-dimensional)
$\theta = g(\mu)$ (e.g., a ratio of two components of μ), $\widehat{\theta} = g(\bar{Y})$
$\mathrm{Var}(\widehat{\theta}) \approx [\nabla g(\mu)]^T \mathrm{Var}(\bar{Y}) \nabla g(\mu)$
An estimator of Var($\widehat{\theta}$) is $\widehat{V}_n = [\nabla g(\bar{Y})]^T (S^2/n) \nabla g(\bar{Y})$

By the SLLN, $\bar{X} \to_{a.s.} \mu$ and $S^2 \to_{a.s.} \mathrm{Var}(X_1)$.

Hence, $\widehat{V}_n$ is strongly consistent for $\mathrm{Var}(\widehat{\theta}_n)$, provided that $\nabla g(\mu) \neq 0$ and ∇g is continuous at μ.

Example 5.15

Let $Y_1, ..., Y_n$ be i.i.d. random variables with finite $\mu_y = EY_1$, $\sigma_y^2 = \mathrm{Var}(Y_1)$, $\gamma_y = EY_1^3$, and $\kappa_y = EY_1^4$.
Consider the estimation of $\theta = (\mu_y, \sigma_y^2)$.
Let $\widehat{\theta}_n = (\bar{X}, \widehat{\sigma}_y^2)$, where $\widehat{\sigma}_y^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$.
If $X_i = (Y_i, Y_i^2)$, then $\widehat{\theta}_n = g(\bar{X})$ with $g(x) = (x_1, x_2 - x_1^2)$

$$\mathrm{Var}(X_1) = \begin{pmatrix} \sigma_y^2 & \gamma_y - \mu_y(\sigma_y^2 + \mu_y^2) \\ \gamma_y - \mu_y(\sigma_y^2 + \mu_y^2) & \kappa_y - (\sigma_y^2 + \mu_y^2)^2 \end{pmatrix}$$

and

$$\nabla g(x) = \begin{pmatrix} 1 & 0 \\ -2x_1 & 1 \end{pmatrix}.$$

The estimator $\widehat{V}_n$ is strongly consistent, since $\nabla g(x)$ is continuous.

Is the derivative ∇g always easy to derive?

An alternative method?

Suppose we can independently obtain B copies of the data set X
Say $X^1,...,X^B$
Then we can calculate $\widehat{\theta}^b = \widehat{\theta}(X^b)$, $b = 1,...,B$
Variance of $\widehat{\theta}$ can be estimated as

$$\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}^b - \frac{1}{B}\sum_{l=1}^{B}\widehat{\theta}^l\right)^2$$

No derivation is needed (at the expense of more computations)

These estimators are valid for large B ($B \to \infty$, law of large numbers).

But typically, we only have one dataset, X.

Pseudo-replications

Can we apply the same idea by creating B pseudo-replicate datasets?

This means $X^1,...,X^B$ are "copies" of X, but they are not independent of X (in fact, they are dependent on X)

These methods are called *resampling* or *replication* methods.

Jackknife

Consider pseudo replicates $X^i = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$, $i = 1, ..., n$.
Let $\widehat{\theta}_{-i}$ be the same estimator as $\widehat{\theta}_n$ but based on X^i, $i = 1, ..., n$.
Since $\widehat{\theta}_n$ and $\widehat{\theta}_{-1}, ..., \widehat{\theta}_{-n}$ estimate the same quantity, the following "sample variance" can be used as a measure of the variation of $\widehat{\theta}_n$:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{\theta}_{-i} - \bar{\theta}_n\right)^2, \qquad \bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{-i}$$

Two issues:

- $\widehat{\theta}_{-i}$'s are not independent.
- $\widehat{\theta}_{-i} - \widehat{\theta}_{-j}$ usually converges to 0 at a fast rate (such as n^{-1}).

If $\widehat{\theta}_n = \bar{X}$ is the sample mean, then $\widehat{\theta}_{-i} - \bar{\theta}_n = (n-1)^{-1}(\bar{X} - X_i)$ and

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{\theta}_{-i} - \bar{\theta}_n\right)^2 = \frac{1}{(n-1)^3}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{S^2}{(n-1)^2}$$

Thus, the correction factor $(n-1)^2/n$ should be multiplied, which leads to the *jackknife variance estimator* of $\mathrm{Var}(\widehat{\theta}_n)$

$$\widehat{V}_J = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{-i} - \bar{\theta}_n\right)^2.$$

Theorem 5.17.

Let $X_1, ..., X_n$ be i.i.d. random d-vectors from F with finite $\mu = E(X_1)$ and $\mathrm{Var}(X_1)$, and let $\widehat{\theta}_n = g(\bar{X})$. Suppose that ∇g is continuous at μ and $\nabla g(\mu) \neq 0$. Then the jackknife variance estimator $\widehat{V}_J$ is strongly consistent for $\mathrm{Var}(\widehat{\theta}_n)$.

Proof.

Let $\bar{X}_{-i}$ be the sample mean based on $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$.
From the mean-value theorem, we have

$$\begin{aligned}
\widehat{\theta}_{-i} - \widehat{\theta}_n &= g(\bar{X}_{-i}) - g(\bar{X}) \\
&= [\nabla g(\xi_{n,i})]^\tau (\bar{X}_{-i} - \bar{X}) \\
&= [\nabla g(\bar{X})]^\tau (\bar{X}_{-i} - \bar{X}) + R_{n,i},
\end{aligned}$$

where $R_{n,i} = \left[\nabla g(\xi_{n,i}) - \nabla g(\bar{X})\right]^\tau (\bar{X}_{-i} - \bar{X})$ and $\xi_{n,i}$ is a point on the line segment between $\bar{X}_{-i}$ and $\bar{X}$.
From $\bar{X}_{-i} - \bar{X} = (n-1)^{-1}(\bar{X} - X_i)$, it follows that $\sum_{i=1}^n (\bar{X}_{-i} - \bar{X}) = 0$ and

$$\frac{1}{n}\sum_{i=1}^n (\widehat{\theta}_{-i} - \widehat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n R_{n,i} = \bar{R}_n.$$

From the definition of the jackknife estimator,

$$\widehat{V}_J = A_n + B_n + 2C_n,$$

where

$$A_n = \frac{n-1}{n}[\nabla g(\bar{X})]^\tau \sum_{i=1}^n (\bar{X}_{-i} - \bar{X})(\bar{X}_{-i} - \bar{X})^\tau \nabla g(\bar{X}),$$

$$B_n = \frac{n-1}{n} \sum_{i=1}^n (R_{n,i} - \bar{R}_n)^2,$$

$$C_n = \frac{n-1}{n} \sum_{i=1}^n (R_{n,i} - \bar{R}_n)[\nabla g(\bar{X})]^\tau (\bar{X}_{-i} - \bar{X}).$$

By $\bar{X}_{-i} - \bar{X} = (n-1)^{-1}(\bar{X} - X_i)$, the SLLN, and the continuity of ∇g at μ,

$$A_n / \operatorname{Var}(\widehat{\theta}_n) \to_{a.s.} 1.$$

Also,

$$(n-1)\sum_{i=1}^n \|\bar{X}_{-i} - \bar{X}\|^2 = \frac{1}{n-1}\sum_{i=1}^n \|X_i - \bar{X}\|^2 = O(1) \text{ a.s.} \qquad (1)$$

Hence $\max_{i\leq n}\|\bar{X}_{-i}-\bar{X}\|^2 \to_{a.s.} 0$, which, together with the continuity of ∇g at μ and $\|\xi_{n,i}-\bar{X}\| \leq \|\bar{X}_{-i}-\bar{X}\|$, implies that

$$u_n = \max_{i\leq n}\|\nabla g(\xi_{n,i}) - \nabla g(\bar{X})\| \to_{a.s.} 0.$$

From (1), $\sum_{i=1}^n \|\bar{X}_{-i}-\bar{X}\|^2/\operatorname{Var}(\widehat{\theta}_n) = O(1)$ a.s. and

$$\frac{B_n}{\operatorname{Var}(\widehat{\theta}_n)} \leq \frac{n-1}{\operatorname{Var}(\widehat{\theta}_n)n}\sum_{i=1}^n R_{n,i}^2 \leq \frac{u_n}{\operatorname{Var}(\widehat{\theta}_n)}\sum_{i=1}^n \|\bar{X}_{-i}-\bar{X}\|^2 \to_{a.s.} 0.$$

By the Cauchy-Schwarz inequality,
$(C_n/V_n)^2 \leq (A_n/V_n)(B_n/V_n) \to_{a.s.} 0$, which completes the proof.

Discussion

- A key step in the proof is that $\widehat{\theta}_{-i} - \widehat{\theta}_n$ can be approximated by $[\nabla g(\bar{X})]^\tau(\bar{X}_{-i}-\bar{X})$, which indicates that $\widehat{V}_J$ is consistent for $\widehat{\theta}_n$ that can be well approximated by some linear statistic.
- More results can be found in Shao and Tu (1995, Chapter 2).
- The jackknife method can be applied to non-i.i.d. problems.

Bootstrap

Create bootstrap pseudo-replicate datasets $X^{*1}, ..., X^{*B}$ randomly generated from X.

Let $\widehat{\theta}^{*b}$ be the same as an estimator $\widehat{\theta}$ but based on X^{*b}, $b = 1, ..., B$.

Is

$$\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}^{*b} - \frac{1}{B}\sum_{l=1}^{B}\widehat{\theta}^{*l}\right)^2 \tag{2}$$

still a valid estimator of $\mathrm{Var}(\widehat{\theta})$?

In fact, the cdf $G(t) = P(\widehat{\theta} - \theta \leq t)$ can be estimated as

$$\frac{1}{B}\sum_{b=1}^{B} I\left(\widehat{\theta}^{*b} - \widehat{\theta} \leq t\right) = \frac{\#\text{ of } b\text{'s such that } \widehat{\theta}^{*b} - \widehat{\theta} \leq t}{B} \tag{3}$$

where $I(A)$ is the indicator function of A.

The answer to this question depends on

- how the sample X is taken
- how $X^1, ..., X^B$ are constructed
- the type of the estimator, $\widehat{\theta}$

A heuristic description for the bootstrap

$\mathscr{P}$: the population producing data X
$\widehat{\mathscr{P}}$: an estimated of the population based on data X
X^*: the bootstrap data produced by $\widehat{\mathscr{P}}$

$$\begin{array}{lccccc} \text{real world} & \mathscr{P} & \Rightarrow & X & \Rightarrow & \widehat{\theta} = \widehat{\theta}(X) \\ \text{bootstrap} & \widehat{\mathscr{P}} & \Rightarrow & X^* & \Rightarrow & \widehat{\theta}^* = \widehat{\theta}(X^*) \end{array}$$

$\text{Var}(\widehat{\theta})$ and $G(t) = P(\widehat{\theta} - \theta \leq t)$ can be approximated by $\mathrm{Var}_*(\widehat{\theta}^*)$ and $\widehat{G}(t) = P_*(\widehat{\theta}^* - \widehat{\theta} \leq t)$, respectively, where the variance and probability are taken under the bootstrap sampling conditioned on X.

If $\widehat{\mathscr{P}}$ is close to $\mathscr{P}$, then

- $\widehat{G}(t)$ is close to $G(t)$;
- $\text{Var}_*(\widehat{\theta}^*)$ is close to $\text{Var}(\widehat{\theta})$.

Note that $\text{Var}_*(\widehat{\theta}^*)$ and $\widehat{G}(t)$ are functions of X and are estimators.
If they have explicit forms, then they can be directly used.
If not, then we approximate them by the Monte Carlo approximations (2) and (3), respectively, based on bootstrap data sets $X^{*1}, ..., X^{*B}$ (copies of X^*).

How do we generate X^* based on X?

Parametric bootstrap

Let $X_1, ..., X_n$ be iid with a cdf F_θ where θ is an unknown parameter vector and F_θ is known when θ is known.
Let $\widehat{\theta}$ be an estimator of θ based on $X = (X_1, ..., X_n)$.
Parametric bootstrap data set $X^* = (X_1^*, ..., X_n^*)$ is obtained by generate iid $X_1^*, ..., X_n^*$ from $F_{\widehat{\theta}}$.

Example: location-scale problems

Let $F_\theta(x) = F_0\left(\frac{x-\mu}{\sigma}\right)$, where $\mu = E(X_1)$, $\sigma^2 =$Var(X_1) and F_0 is a known cdf.
Let $\bar{X}$ be the sample mean, S^2 be the sample variance, and

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} = \sqrt{n}\sum_{i=1}^{n} \frac{X_i - \mu}{S}$$

The distribution of T does not depend on any parameter.
It is the t-distribution with degrees of freedom $n-1$ if F_0 is normal.
Otherwise its explicit form is unknown.

Let $\widehat{\theta} = (\bar{X}, S^2)$ Generate iid X_i^*, $i = 1, ..., n$, from $F_{\widehat{\theta}}$.
Then $(X_i^* - \bar{X})/S^2 \sim F_0$

$$T^* = \sqrt{n} \sum_{i=1}^{n} \frac{X_i^{*b} - \bar{Y}}{S} \sim T$$

The parametric bootstrap is prefect: $\mathrm{Var}^*(T^*) = \mathrm{Var}(T)$.
If we calculate $\mathrm{Var}^*(T^*)$ by Monte Carlo approximation, then the parametric bootstrap is exactly the same as the simulation approach.
In general, if there is a function τ such that

$$\mathrm{Var}_\theta(\widehat{\theta}) = \tau(\theta), \quad X_1, ..., X_n \text{ are iid from } F_\theta$$

then

$$\mathrm{Var}^*_{\widehat{\theta}}(\widehat{\theta}^*) = \tau(\widehat{\theta}), \quad X_1^*, ..., X_n^* \text{ are iid from } F_{\widehat{\theta}}$$

Hence, the parametric bootstrap is simply the substitution approach.
If $\widehat{\theta}$ is consistent and τ is continuous, then $\mathrm{Var}^*_{\widehat{\theta}}(\widehat{\theta}^*)$ is consistent.
If τ does not have a close form, we apply Monte Carlo approximation.
In the location-scale example, $\tau =$ a constant and hence the bootstrap is prefect.

Example

Let $X_1,...,X_n$ be iid from F_θ.
Define $\mu = \mu(\theta) = E_\theta(X_1)$, $\mu_j = \mu_j(\theta) = E_\theta(X_1 - \mu)^j$, $j = 2,3,4$.
Consider the estimation of μ^2 by $\bar{X}^2$.
A direct calculation shows that

$$\mathrm{Var}_\theta(\bar{X}^2) = \frac{4[\mu(\theta)]^2\mu_2(\theta)}{n} + \frac{4\mu(\theta)\mu_3(\theta)}{n^2} + \frac{\mu_4(\theta)}{n^3}$$

Based on the previous discussion, the parametric bootstrap variance estimator is

$$\mathrm{Var}^*_\theta(\bar{X}^{*2}) = \frac{4[\mu(\widehat{\theta})]^2\mu_2(\widehat{\theta})}{n} + \frac{4\mu(\widehat{\theta})\mu_3(\widehat{\theta})}{n^2} + \frac{\mu_4(\widehat{\theta})}{n^3}$$

It is a consistent estimator if μ, μ_j, $j = 2,3,4$, are continuous functions.

If we apply the asymptotic approach, then we estimate $\mathrm{Var}_\theta(\bar{X}^2)$ by

$$\frac{4[\mu(\widehat{\theta})]^2\mu_2(\widehat{\theta})}{n}$$

Nonparametric bootstrap

Without any model, we can apply the simple nonparametric bootstrap.
If $X = (X_1, ..., X_n)$, $X_1, ..., X_n$ are iid, then $\mathscr{P}$ is the cdf of X_1 and $\widehat{\mathscr{P}}$ is the empirical cdf based on $X_1, ..., X_n$.
If we generate iid bootstrap data $X_1^*, ..., X_n^*$ from $\widehat{\mathscr{P}}$, then it is the same as taking a simple random sample with replacement from X.

Property of $\mathrm{Var}^*(\widehat{\theta}^*)$

Consider first $\widehat{\theta} = \bar{X}$, the sample mean, $\widehat{\theta}^* = \bar{X}^*$, the sample mean of $X_1^*, ..., X_n^*$.

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

$$\begin{aligned}
\mathrm{Var}^*(\bar{X}^*) &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}^*(X_i^*) = \frac{1}{n^2}\sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})^\tau \\
&= \frac{1}{n^2}\sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})^\tau = \frac{n-1}{n^2} S^2 \approx \frac{S^2}{n}
\end{aligned}$$

When n is small, we may make an adjustment of $\frac{n}{n-1}$.

Consider next the estimation of $g(\mu)$, where $\mu = E(X_1)$ and g is a continuously differentiable function.
Our estimator is $\widehat{\theta} = g(\bar{X})$.
The bootstrap analog is $\widehat{\theta}^* = g(\bar{X}^*)$.
When n is large,

$$g(\bar{X}^*) = g(\bar{X}) + \nabla g(\bar{X})(\bar{X}^* - \bar{X}) + \cdots \approx g(\bar{X}) + \nabla g(\bar{X})(\bar{X}^* - \bar{X})$$

and

$$\begin{aligned} \mathrm{Var}^*(\widehat{\theta}^*) &= \mathrm{Var}^*[g(\bar{X}^*)] \\ &\approx \nabla g(\bar{X}) \mathrm{Var}^*(\bar{X}^* - \bar{X}) \nabla g^\tau(\bar{X}) \\ &\approx \frac{n-1}{n^2} \nabla g(\bar{X}) S^2 \nabla g^\tau(\bar{X}) \end{aligned}$$

Example

Let $X_1, ..., X_n$ be iid from F.
Define $\mu = E_\theta(X_1)$, $\mu_j = E_\theta(X_1 - \mu)^j$, $j = 2, 3, 4$.
Consider the estimation of μ^2 by $\bar{X}^2$:

$$\mathrm{Var}(\bar{X}^2) = \frac{4\mu^2\mu_2}{n} + \frac{4\mu\mu_3}{n^2} + \frac{\mu_4}{n^3}$$

and

$$\mathrm{Var}^*(\bar{X}^{*2}) = \frac{4\bar{X}^2 m_2}{n} + \frac{4\bar{X} m_3}{n^2} + \frac{m_4}{n^3}$$

where

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^j, \quad j = 2,3,4.$$

This is because the mean of the empirical cdf $\widehat{F}_n$ is $\bar{X}$ and the jth central moment of $\widehat{F}_n$ is is m_j.

In this case, we have an explicit form for the bootstrap variance estimator $\mathrm{Var}^*(\bar{X}^{*2})$ so no Monte Carlo is needed.

This bootstrap variance estimator is consistent, since sample moments m_j's are consistent for μ_j's, by the WLLN.

Since $g'(x) = 2x$ when $g(x) = x^2$, the use of the approximation derived eariler gives that

$$\mathrm{Var}^*(\bar{X}^{*2}) \approx \frac{4\bar{X}^2 m_2}{n}$$

which is also consistent since the terms ignored are of the orders n^{-2}.

In fact, the delta-method produces the variance estimator

$$\frac{[g'(\bar{X})]^2 S^2}{n} = \frac{4\bar{X}^2 S^2}{n}$$

Discussion

- In general, the expression $\mathrm{Var}^*(\widehat{\theta}^*)$ is usually complicated and not explicit, so Monte Carlo approximation is necessary.
- In fact, the idea of using the bootstrap is not to derive its explicit form (since it involves complex derivations).
- The bootstrap is to replace theoretical derivations by repeated computations.
- In many cases the theoretical derivations are difficult or messy.
- The user does not need to do theoretical derivations.
- However, they should be told when using the bootstrap produces correct variance estimators and how to do the bootstrap.
- The research on the bootstrap methodology still requires theoretical derivations.
- The jackknife shares the same idea as the bootstrap in some aspects, e.g., replacing theoretical derivations by repeated computations, but the bootstrap is a more complete methodology.