Bootstrap for complex survey data

Stas Kolenikov

Department of Statistics
University of Missouri-Columbia

JSM 2009
Washington, DC

Educational objectives

Upon completion of this course, you will

- become familiar with main variance estimation methods for complex survey data, their strengths and weaknesses
- be able to identify appropriate variance estimation methods depending on the sample design, complexity of the problem, confidentiality protection
- know how to utilize the existing bootstrap weights
- know how to create bootstrap weights in Stata and R
- know how to choose parameters of the bootstrap

Outline

The bootstrap for i.i.d. data

1. Bootstrap principle
2. Bootstrap bias and variance estimates
3. Bootstrap confidence intervals
4. More bootstrap theory
5. Some extensions

Bootstrap principle

- Population: distribution F, parameter $\theta = T(F)$, both can be multivariate
- Sample: data $X_1, \ldots, X_n \sim$ i.i.d. F, distribution F_n, parameter estimate $\hat{\theta}_n = T(F_n)$
- Inference: need to know distribution $\mathcal{D}[\hat{\theta}_n]$, often in asymptotic form $\mathbf{Pr}[\sqrt{n}(\hat{\theta}_n - \theta) < x]$
- Bootstrap: use F_n to take samples from
- Bootstrap samples: $X_1^*, \ldots, X_n^* \sim$ i.i.d. F_n, distribution F_n^*, parameter estimate $\hat{\theta}_n^* = T(F_n^*)$

$$\begin{array}{ccccc} F & \overset{\text{sample}}{\longrightarrow} & F_n & \overset{\text{bootstrap}}{\longrightarrow} & F_n^* \\ \downarrow T & & \downarrow T & & \downarrow T \\ \theta & \overset{?}{\longleftrightarrow} & \hat{\theta}_n & \overset{\text{bootstrap}}{\longleftrightarrow} & \hat{\theta}_n^* \end{array}$$

Aside: what is T?

T does something to distribution F that results in a number or a vector: $\theta = T(F)$.

- T finds a point where $F(x) = 1/2$: θ is the median of the distribution
- T takes an expected value with respect to F:

$$\theta = \mathbb{E}[X] = \int xF(dx)$$

- T finds a solution to $\int (y - \theta x) F(dx, dy) = 0$:

$$\theta = \mathbb{E}[y] / \mathbb{E}[x]$$

Bootstrap principle

$$\begin{array}{ccccc} F & \xrightarrow{\text{sample}} & F_n & \xrightarrow{\text{bootstrap}} & F_n^* \\ \downarrow T & & \downarrow T & & \downarrow T \\ \theta & \overset{?}{\longleftrightarrow} & \hat\theta_n & \overset{\text{bootstrap}}{\longleftrightarrow} & \hat\theta_n^* \end{array}$$

Theoretical/ideal/complete bootstrap: sampling distributions over all n^n possible samples

$$\begin{aligned} \text{Bias}[\hat\theta_n] &= \mathbb{E}[\hat\theta_n - \theta] &&\doteq \mathbb{E}^*[\hat\theta_n^* - \hat\theta_n|\mathbf{X}] \\ \mathbb{V}[\hat\theta_n] &= \mathbb{E}\left[(\hat\theta_n - \mathbb{E}[\hat\theta_n])^2\right] &&\doteq \mathbb{E}^*\left[(\hat\theta_n^* - \mathbb{E}[\hat\theta_n^*])^2|\mathbf{X}\right] \\ \text{MSE}[\hat\theta_n] &= \mathbb{E}\left[(\hat\theta_n - \theta)^2\right] &&\doteq \mathbb{E}^*\left[(\hat\theta_n^* - \theta_n])^2|\mathbf{X}\right] \\ F_{\theta_n - \theta}(x) &= \mathbf{Pr}[\hat\theta_n - \theta < x] &&\doteq \mathbf{Pr}^*[\hat\theta_n^* - \hat\theta_n < x|\mathbf{X}] \end{aligned} \quad (1)$$

Monte Carlo bootstrap

As taking n^n bootstrap samples is not feasible, use Monte Carlo simulation instead:

1. For the r-th bootstrap sample, take a simple random sample with replacement $X_1^{(*r)}, \ldots, X_n^{(*r)}$ from $X_1, \ldots, X_n$.
2. Compute the parameter estimate of interest $\hat{\theta}_n^{(*r)}$.
3. Repeat Steps 1–2 for $r = 1, \ldots, R$.
4. Approximate the ideal bootstrap distribution with distribution of $\hat{\theta}_n^{(*1)}, \ldots, \hat{\theta}_n^{(*R)}$.

$$
\begin{array}{ccccccc}
F & \xrightarrow{\text{sample}} & F_n & \xrightarrow{\text{bootstrap}} & [F_n^*] & \xrightarrow{\text{simulate}} & F_n^{(*r)} \\
\downarrow T & & \downarrow T & & \downarrow T & & \downarrow T \\
\theta & \overset{?}{\longleftrightarrow} & \hat{\theta}_n & \xleftrightarrow{\text{bootstrap}} & [\hat{\theta}_n^*] & \approx & \hat{\theta}_n^{(*r)}
\end{array}
$$

Estimates of bias and variance

Estimate of the bias:

$$\text{Bias}[\hat\theta_n] = \mathbb{E}[\hat\theta_n - \theta] \doteq \mathbb{E}^*[\hat\theta_n^* - \hat\theta_n | \mathbf{X}] \approx \frac{1}{R}\sum_{r=1}^{R} \hat\theta_n^{(*r)} - \hat\theta_n$$

Bias corrected estimate:

$$\tilde\theta_n = 2\hat\theta_n - \frac{1}{R}\sum_{r=1}^{R} \hat\theta_n^{(*r)}$$

Variance estimate:

$$\mathbb{V}[\hat\theta_n] = \mathbb{E}\left[(\hat\theta_n - \mathbb{E}[\hat\theta_n])^2\right] \doteq \mathbb{E}^*\left[(\hat\theta_n^* - \mathbb{E}[\hat\theta_n^*])^2 | \mathbf{X}\right]$$

$$\approx \frac{1}{R}\sum_{r=1}^{R}\left(\hat\theta_n^{(*r)} - \frac{1}{R}\sum_{l=1}^{R}\hat\theta_n^{(*l)}\right)^2 \equiv v_{BOOT}[\hat\theta_n]$$

Number of samples

How to chose the number of the bootstrap samples R?

- Stability of the standard errors:

$$\mathrm{cv}^*(s_R) \approx \sqrt{\frac{\hat\kappa + 2}{4R}}$$

  where $\hat\kappa$ is the kurtosis of $\hat\theta_n^*$
- Confidence interval accuracy:

$$\mathrm{cv}^*(CB_R - \hat\theta_n) \approx \frac{1}{|z_\alpha|}\sqrt{\frac{1}{R}\left(\frac{1}{\phi(0)^2} - \frac{2(1-\alpha)}{\phi(0)\phi(z_\alpha)} + \frac{\alpha(1-\alpha)}{\phi(z_\alpha)^2}\right)}$$

  where CB_R is the confidence bound with level $1-\alpha$
- Estimation of moments: $R = 50$–200
- Estimation of quantiles/distribution functions: $R \geq 1000$

Percentile confidence intervals

Idea:

$$\mathbf{Pr}[\hat{\theta}_n - \theta < x] \doteq \mathbf{Pr}^*[\hat{\theta}^*_n - \hat{\theta}_n < x|\mathbf{X}]$$

Lower confidence bound of level α:

$$K^{-1}_{BOOT}(\alpha)$$

where

$$K_{BOOT}(x) = \mathbf{Pr}^*[\hat{\theta}^*_n \leq x]$$

is the (ideal or Monte Carlo) bootstrap distribution of $\hat{\theta}^*_n$.

Normal confidence intervals

Idea:

$$\hat{\theta}_n \approx N(\theta, \sigma_n^2) \quad \doteq \quad \hat{\theta}_n^* \approx N(\hat{\theta}, \sigma_n^{*2})$$

Lower confidence bound of level α:

$$\hat{\theta}_n + \sigma_n^* \Phi^{-1}(\alpha)$$

where σ_n^{*2} is the variance of the bootstrap distribution.

Bootstrap-t CI

Idea: pivotal quantity

$$t = (\hat{\theta}_n - \theta)/\hat{\sigma}_n$$

has asymptotic distribution that is the same for all $\langle F, \theta \rangle$.

Lower confidence bound of level α:

$$\hat{\theta}_n - \hat{\sigma}_n G^{-1}_{BOOT}(1-\alpha)$$

where

$$G_{BOOT}(x) = \mathbf{Pr}^*[(\hat{\theta}^*_n - \hat{\theta}_n)/\hat{\sigma}^*_n \le x]$$

is the bootstrap distribution of the above pivot.

Bias corrected CI

Idea: $\phi_n(\cdot)$ is an increasing transformation (e.g., variance stabilizing, skewness reducing); assume

$$\mathbf{Pr}[\phi_n(\hat{\theta}_n) - \phi_n(\theta) + z_0 \leq x] \approx \Phi(x)$$

Lower confidence bound of level α:

$$K^{-1}_{BOOT}(\Phi(z_\alpha + 2\Phi^{-1}(K_{BOOT}(\hat{\theta}_n))))$$

Accelerated bias corrected CI

Idea:

$$\mathbf{Pr}\left[\frac{\phi_n(\hat{\theta}_n) - \phi_n(\theta)}{1 + a\phi_n(\theta)} + z_0 \leq x\right] \approx \Phi(x)$$

with tuning parameter a correcting for skewness of $\phi_n(\hat{\theta})$.

Lower confidence bound of level α:

$$K^{-1}_{BOOT}(\Phi(z_0 + (z_\alpha + z_0)/(1 - a(z_\alpha + z_0))))$$

Parameter a needs to be computed or estimated, e.g. via the jackknife.

Asymptotic justification of the bootstrap

Let us look at the diagram again:

$$\begin{array}{ccccc} F & \xrightarrow{\text{sample}} & F_n & \xrightarrow{\text{bootstrap}} & F_n^* \\ \downarrow T & & \downarrow T & & \downarrow T \\ \theta & \overset{?}{\longleftrightarrow} & \hat\theta_n & \overset{\text{bootstrap}}{\longleftrightarrow} & \hat\theta_n^* \end{array}$$

When would the relation between $\hat\theta_n \longleftrightarrow \hat\theta_n^*$ be similar to the one between $\hat\theta \longleftrightarrow \hat\theta_n$?

Asymptotic justification of the bootstrap

- The bootstrap can only be successful if F_n is sufficiently close to F for the bootstrap distribution $\mathcal{D}^*[\hat{\theta}_n^*]$ to resemble the sampling distribution $\mathcal{D}[\hat{\theta}_n]$.
- Small deviations of F_n from F must translate to small deviations of $\mathcal{D}^*[\hat{\theta}_n^*]$ from $\mathcal{D}[\hat{\theta}_n]$.
- Taylor series expansion/the delta method for $\theta = T(F)$:

$$\hat{\theta}_n - \theta = \nabla T\big|_F (F_n - F) + o(\|F_n - F\|),$$
$$\hat{\theta}_n^* - \hat{\theta}_n = \nabla T\big|_{F_n} (F_n^* - F_n) + o(\|F_n^* - F_n\|)$$

- Functional T must satisfy some smoothness conditions, and its "derivative" should be bounded away from zero.
- F_n^* must converge to F_n at the same rate as F_n converges to F.

Bootstrap failures

Sometimes, the simple bootstrap as described above produces a misleading answer.

- Non-i.i.d. data: time series, spatial data, clustered surveys, overdispersed count data (Canty, Davison, Hinkley & Ventura 2006)
- Non-regular problems (Shao & Tu 1995, Sec. 3.6)
  - Certain heavy tailed distributions (Canty, Davison, Hinkley & Ventura 2006)
  - Zero derivatives (Andrews 2007): $\bar{X}_n^2$ when $\mu = 0$
  - Non-smooth functions (Bickel & Freedman 1981): $|\bar{X}_n|$, sample quantiles/extreme order statistics/min/max
  - Different rates of convergence (Canty, Davison, Hinkley & Ventura 2006): sample mode, shrinkage and kernel estimators
  - Constrained estimation (Andrews 2000): $\bar{X}_n$ when $\mu \geq 0$

Bootstrap tests

$$H_0 : T(F) = \theta_0 \quad \text{vs.} \quad H_1 : T(F) \neq \theta_0$$

To compute the p^*-values of the bootstrap distribution, one needs to sample from the distribution that satisfies H_0. For continuous problems, the data distribution won't satisfy H_0 with probability 1. The data need to be transformed prior to the bootstrap:

- shift?
- scale?
- rotation?
- reweighting?

Non-parametric flavor will likely be lost.

Balanced bootstrap

Motivation: if $\hat{\theta}_n = T(F_n) = \bar{X}_n$, the complete bootstrap gives $\mathbb{E}^*[\hat{\theta}^*_n] = \bar{X}_n$ and $\mathbb{V}^*[\hat{\theta}^*_n] = s^2/n$. Is it possible to match the moments of the simulated bootstrap?

- Equality for the mean:

$$\frac{1}{R}\sum_r \bar{x}_n^{(*r)} = \frac{1}{nR}\sum_i x_i \sum_r f_i^{(*r)} = \frac{1}{n}\sum_i x_i$$

  $f_i^{(*r)} = \#$ times unit i is used in the r-th bootstrap sample

- First order balance (Davison, Hinkley & Schechtman 1986):

$$\sum_r f_i^{(*r)} = R \text{ for all } i$$

- Practical implementation: permutation of $\{1, \ldots, n\}^R$ (Gleason 1988)

Balanced bootstrap

- Equality for the variance:

$$\frac{1}{Rn^2}\sum_r\Big[\sum_i (f_i^{(*r)}-1)^2 X_i + \sum_{i\neq j}(f_i^{(*r)}-1)(f_j^{(*r)}-1)X_iX_j\Big]$$
$$= \frac{n-1}{n^3}\sum(1-\frac{1}{n})^2 X_i + \frac{1}{n^3}\sum_{i\neq j} X_iX_j$$

- Second order balance (Graham, Hinkley, John & Shi 1990): for all i, j

$$n\sum_r f_i^{(*r)2} = R(2n-1), \quad n\sum_r f_i^{(*r)} f_j^{(*r)} = R(n-1)$$

- Additional restriction: R must be a multiple of n
- Practical implementation: orthogonal arrays and incomplete block designs

Wild bootstrap

Special situation: heteroskedastic regression (Wu 1986) or non-parametric regression (Härdle 1990).

1. Fit regression model $y_i = \widehat{f}(x_i) + e_i$
2. Bootstrap distribution of residuals ϵ_i^* in observation i:

$$\mathbb{E}^*[\epsilon_i^*] = 0, \quad \mathbb{E}^*[\epsilon_i^{*2}] = e_i^2, \quad \mathbb{E}^*[\epsilon_i^{*3}] = e_i^3$$

   Example: two-point golden rule distribution:

$$\epsilon_i^* = e_i(1 \pm \sqrt{5})/2 \text{ with prob. } (5 \mp \sqrt{5})/10$$

3. Form bootstrap samples as $y_i^* = \widehat{f}(x_i) + \epsilon_i^*$

Review questions

1. Explain how the bootstrap can be used to estimate $\mathrm{CV}[\bar{X}_n]$.
2. Suggest a method to compute $\hat{\sigma}_n$ for bootstrap-t confidence interval method.
3. Given that the kurtosis of the bootstrap distribution is 0.5, find the number of replicates needed to make the CV of the bootstrap standard errors equal to 5%.
4. (requires calculus) Assuming all X_i's are distinct, find $\lim_{n\to\infty} \mathbf{Pr}^*[X^*_{(n)} = X_{(n)}]$ where $X_{(n)}$ is the maximum in the data, and $X^*_{(n)}$ is the maximum in the bootstrap sample. *Hint:* find the probability of the complement of this event.

Notes

1. $CV(\bar{x}) = (V(\bar{x}))^{1/2} / \bar{x}$

2. $\hat{\sigma}_n =$? LINEARIZATION, JACKKNIFE, BOOT

3. $CV = \sqrt{\frac{K+2}{4R}} = 0{,}05 = \sqrt{\frac{2.5}{4R}}$, $0.0025 = \frac{2.5}{4R}$

$R = \frac{2.5}{0.0025 \cdot 4} = 250$

4. IPL

Variance estimation for complex surveys

1. Features of complex survey data
2. Linearization variance estimation
3. Replication methods: overview
4. Jackknife
5. BRR

Survey settings

- Complex survey designs include stratification, cluster samples, multiple stages of selection, unequal probabilities of selection, non-response and post-stratification adjustments, longitudinal and rotation features.
- Unless utmost precision is required (or sampling fractions are large), it suffices to approximate real designs by two-stage stratified designs with PSUs sampled with replacement.
- Notation:
  - L = # strata
  - n_h = # units in stratum h
  - PSUs are indexed by i
  - SSUs are indexed by j
  - generic datum is x_{hij}

Variance estimation goals

- Reporting and analytic purposes: a survey analyst needs standard errors to include in the report; an applied researcher needs standard errors to test their substantive models.
- Design purposes: a sample designer needs to know population variances to find efficient designs, strata allocations, small area estimators.

Explicit variance formulae

For a (very) limited number of statistics, explicit variance formulae are available.

Horvitz-Thompson estimator:

$$t_{HT}[x] = \sum_{i \in \mathcal{S}} \frac{x_i}{\pi_i}$$

Design variance:

$$\mathbb{V}\left[t_{HT}[x]\right] = \frac{1}{2} \sum_{i \neq j \in \mathcal{U}} (\pi_i \pi_j - \pi_{ij}) \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j} \right)^2$$

Yates-Grundy-Sen variance estimator:

$$v_{YGS} = \frac{1}{2} \sum_{i \neq j \in \mathcal{S}} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j} \right)^2$$

Explicit variance formulae

Stratified sample:

$$v_{str}\left[t_{str}[x]\right] = \sum_{h=1}^{L}(1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(t_{hi}-\bar{t}_h)^2$$

$$t_{hi} = \sum_{j\in \mathrm{PSU}_{hi}} \frac{x_{hij}}{\pi_{hij}}$$

$$\bar{t}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} t_{hi}$$

Linearization variance estimator

- $\theta = f(T[x_1], \ldots, T[x_k])$ is a function of moments
- $\hat{\theta} = f(t[x_1], \ldots, t[x_k])$ is its estimator
- Taylor series expansion/delta method:

$$\hat{\theta} = \theta + \nabla f(t[\mathbf{x}] - T[\mathbf{x}]) + \ldots$$

- Hence

$$v_L[\hat{\theta}] \approx \widehat{\mathrm{MSE}}[\hat{\theta}] \approx v\left[\sum_k \frac{\partial f}{\partial t_k} t_k\right]$$

- Regularity conditions: $\partial f / \partial t_k |_{T[\mathbf{x}]} \neq 0$.
- Example: ratio $r = t[y]/t[x]$, variance estimator

$$v_L[r] = \frac{1}{t[x]^2} v(e_i), \quad e_i = y_i - r x_i, \quad T[x] \neq 0$$

Linearization variance estimator

- $\hat{\theta}$ solves estimating equations

$$g(\mathbf{x}, \hat{\theta}) = \sum_{i \in \mathcal{S}} \frac{g(x_i, \hat{\theta})}{\pi_i} = 0$$

- Taylor series expansion:

$$g(\mathbf{x}, \hat{\theta}) - g(\mathbf{x}, \theta) = \nabla g \cdot (\hat{\theta} - \theta) + \ldots$$

- Invert it and account for $g(\mathbf{x}, \hat{\theta}) = 0$ to obtain

$$\hat{\theta} - \theta = -(\nabla g)^{-1} g(\mathbf{x}, \theta) + \ldots$$

- Take the variance and plug the estimates:

$$v_L[\hat{\theta}] \approx \widehat{\mathrm{MSE}}[\hat{\theta}] \approx (\nabla g)^{-1} v[g(\mathbf{x}, \hat{\theta})] (\nabla g)^{-1T}$$

- Example: GLM (Binder 1983)

Replication methods

For a given estimation procedure $(X_1, \ldots, X_n) \mapsto \hat{\theta}$:

1. To create data for replicate r, reshuffle PSUs, omitting some and/or repeating others, according to a certain replication scheme.
2. Using the original estimation procedure and the replicate data, obtain parameter estimate $\hat{\theta}^{(r)}$.
3. Repeat Steps 1–2 for $r = 1, \ldots, R$.
4. Estimate variance/MSE as

$$v_m[\hat{\theta}] = \frac{A}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \tilde{\theta})^2 \tag{2}$$

where A is a scaling parameter, $\tilde{\theta} = \sum_r \hat{\theta}^{(r)}/R$ for variance estimation and $\tilde{\theta} = \hat{\theta}$ for MSE estimation.

Alternative implementation: replicate weights $w_{hij}^{(r)}$

Pros and cons of resampling estimators

+ Only need software that does weighted estimation; no need to program specific estimators for each model
+ No need to release unit identifiers in public data sets
– Computationally intensive
– Post-stratification and non-response adjustments need to be performed on every set of weights
– Bulky data files with many weight variables

The jackknife

Kish & Frankel (1974), Krewski & Rao (1981)

- Replicates: omit only one PSU from the entire sample
- Replicate weights: if unit k from stratum g is omitted,

$$w_{hij}^{(gk)} = \begin{cases} 0, & h = g, i = k \\ \frac{n_g}{n_g - 1} w_{hij}, & h = g, i \neq k \\ w_{hij}, & h \neq g \end{cases}$$

- Number of replicates: $R = n$
- Scaling factor in (2):

$$A = \begin{cases} n - 1, & L = 1 \\ n_h - 1 \text{ within strata}, & L > 1 \end{cases}$$

The jackknife

Variance estimators:

$$v_{J1} = \sum_h \frac{n_h - 1}{n_h} \sum_i (\hat\theta^{(hi)} - \hat\theta^h)^2$$

$$v_{J2} = \sum_h \frac{n_h - 1}{n_h} \sum_i (\hat\theta^{(hi)} - \hat\theta)^2$$

$$v_{J3} = \sum_h \frac{n_h - 1}{n_h} \sum_i (\hat\theta^{(hi)} - \sum_g \sum_k \hat\theta^{(gk)}/n)^2$$

$$v_{J4} = \sum_h \frac{n_h - 1}{n_h} \sum_i (\hat\theta^{(hi)} - \sum_h \hat\theta^h/L)^2$$

where

$$\hat\theta^h = \sum_i \hat\theta^{(hi)}/n_h$$

The jackknife

Pseudo-values:

$$\tilde{\theta}^{(hi)} = n_h \hat{\theta}^h - (n_h - 1)\hat{\theta}^{(hi)}$$

More variance estimators:

$$v_{J5} = \sum_h \frac{1}{(n_h - 1)n_h} \sum_i (\tilde{\theta}^{(hi)} - \sum_g \sum_k \tilde{\theta}^{(gk)}/n)^2$$

$$v_{J6} = \sum_h \frac{1}{(n_h - 1)n_h} \sum_i (\tilde{\theta}^{(hi)} - 1/L \sum_g 1/n_g \sum_k \tilde{\theta}^{(gk)})^2$$

Bias corrected point estimator:

$$\hat{\theta}_J = (n + 1 - L)\hat{\theta} - \sum_h (n_h - 1)\hat{\theta}^h$$

The jackknife/linearization failures

Linearization and the jackknife estimators are inconsistent for non-smooth parameters:

- Percentiles (including median)
- Extreme order statistics: min, max
- Exotic estimation problems: $|\theta|$, matching estimators

Delete-k jackknife

If $n_h > k > 1$ for all h, a variation of the jackknife is to delete k PSUs at a time rather than one.

- Replicate weight:

$$w_{hij}^{(r)} = \begin{cases} 0, & \text{unit } hi \text{ is omitted,} \\ \dfrac{n_h}{n_h - k} w_{hij}, & \text{units in the same stratum are omitted but not } hi, \\ w_{hij}, & \text{units in stratum other than } h \text{ are omitted} \end{cases}$$

- Number of replicates: $R = \sum_h \binom{n_h}{k}$
- Scaling factor in (2): $(n_h - k)/k$, within strata
- Pros: better performance in non-smooth problems
- Cons: increased computational complexity

Balanced repeated replication (BRR)

- Design restriction: $n_h = 2$ PSUs/stratum
- Replicates (half-samples): omit one of the two PSUs from each stratum
- Replicate weights:

$$w_{hij}^{(r)} = \begin{cases} 2w_{hij}, & \text{PSU } hi \text{ is retained} \\ 0, & \text{PSU } hi \text{ is omitted} \end{cases}$$

- (2nd order) balance conditions:
  - each PSU is used $R/2$ times
  - each pair of PSUs is used $R/4$ times
- Number of replicates: $L \le R \le 2^L$
- McCarthy (1969): $L \le R = 4m \le L + 3$ using Hadamard matrices
- Scaling factor in (2): $A = 1$

Aside: Hadamard matrices

- $n \times n$ matrix with entries ± 1
- Rows are orthogonal
- Special case of orthogonal arrays (Hedayat, Sloane & Stufken 1999)
- Hadamard conjecture: for every integer m, there exists an Hadamard matrix of order $4m$
- Smallest order for which no matrix is known: $4m = 668$
- Sylvester construction for orders 2^k: if H is Hadamard, so is

$$\begin{pmatrix} H & H \\ H & -H \end{pmatrix}$$

- BRR designs: $w_{hi}^{(r)} = (1 + H_{rh}) w_{hi}$

BRR

Complementary half-samples: swap included/excluded units, obtain $\hat\theta^{(rc)}$.
Variance estimators:

$$v_{BRR1}[\hat\theta] \equiv v_{BRR-H}[\hat\theta] = \frac{1}{R}\sum_{r=1}^{R}(\hat\theta^{(r)}_{BRR} - \hat\theta)^2$$

$$v_{BRR2}[\hat\theta] \equiv v_{BRR-D}[\hat\theta] = \frac{1}{4R}\sum_{r=1}^{R}(\hat\theta^{(r)}_{BRR} - \hat\theta^{(rc)}_{BRR})^2$$

$$v_{BRR3}[\hat\theta] \equiv v_{BRR-S}[\hat\theta] = \frac{1}{2R}\sum_{r=1}^{R}(\hat\theta^{(r)}_{BRR} - \tilde\theta)^2 + (\hat\theta^{(rc)}_{BRR} - \tilde\theta)^2$$

Bias corrected estimate:

$$\hat\theta_{Bc} = 2\hat\theta - \frac{1}{R}\sum_r \hat\theta^{(r)} \quad \text{or} \quad 2\hat\theta - \frac{1}{2R}\sum_r (\hat\theta^{(r)} + \hat\theta^{(rc)})$$

Fay's modification

- Confidentiality protection: if units have a replicate weight of 0 they belong to the same PSU
- Modified weights:

$$w_{hi}^{(r)} = (1 + kH_{rh})w_{hi}$$

  for some $0 < k \leq 1$
- Scaling constant in (2): $A = 1/k^2$

Extensions of BRR

What if $n_h \geq 2$?

- Gurney & Jewett (1975): $n_h = p$ for a prime p, $R = (p^k - 1)/(p - 1) \geq L$
- Gupta & Nigam (1987) and Wu (1991): mixed orthogonal arrays for $n_h \geq 2$, 1 PSU/stratum recycled, $R =$?
- Sitter (1993): orthogonal multiarrays for $n_h \geq 2$, about half PSUs/stratum recycled, $R =$?
- Availability of a suitable orthogonal array needs to be established for each particular design

Approximate BRR

Since BRR is a common estimation technique, some publicly released data use design approximations that would allow the end user to use BRR techniques:

- strata collapse
- grouping of PSUs
- treating SSUs as PSUs for self-representing units

Caution: Shao (1996) gives an example where grouped BRR is inconsistent.

Remedies: repeated grouping, random subsampling.

Review questions

1. (requires calculus) If variance estimator $v[\hat{\theta}]$ is available for parameter estimate $\hat{\theta}$, what is $v_L[e^{\hat{\theta}}]$?
2. True or false: In regression analysis, the linear model textbook variance estimator $s^2(X'X)^{-1}$ is appropriate for complex survey data.
3. For a design with 2 PSUs/stratum, which method will be faster, the jackknife or BRR?
4. If $L = 45$ and $n_h = 2$ for every stratum, can one construct BRR designs with $R = 50$? $R = 60$? What's the smallest number of replicates necessary?

Notes

Complex survey bootstraps

1. Naïve bootstrap
2. Rescaling bootstrap
3. Other survey bootstraps:
   - bootstrap without replacement
   - mirror-match bootstrap
   - mean bootstap
   - bootstrap for imputed data
   - balanced bootstrap
   - variance components bootstrap
   - wild bootstrap
   - parametric bootstrap for small area estimation
4. Comparison of all methods

How about some theory?

Naïve bootstrap

1. Sample with replacement n_h units from stratum h.
2. For each replicate, compute $\hat{\theta}^{(r)}$.
3. Estimate the variance using (2).

Rao & Wu (1988):

$$\mathbb{V}^*[\bar{x}^*] = \sum_h \frac{W_h^2}{n_h} \frac{n_h - 1}{n_h} s_h^2$$

rather than

$$v[\bar{x}] = \sum_h \frac{W_h^2}{n_h} s_h^2$$

Scaling issue? Choice of A?

Rescaling bootstrap (RBS)

Rao & Wu (1988): for parameter $\theta = f(\bar{\mathbf{x}})$,

1. Sample with replacement m_h out of n_h units in stratum h.
2. Compute pseudo-values

$$\tilde{x}_h^{(r)} = \bar{x}_h + m_h^{1/2}(n_h - 1)^{-1/2}(\bar{x}_h^{(*r)} - \bar{x}_h),$$
$$\tilde{x}^{(r)} = \sum_h W_h \tilde{x}_h^{(r)}, \quad \tilde{\theta}^{(r)} = f(\tilde{x}^{(r)}) \tag{3}$$

3. Repeat Steps 1–2 for $r = 1, \ldots, R$.
4. Compute $v_{RBS}[\hat{\theta}]$ using (2) with $A = 1$.

Scaling of weights

Rao, Wu & Yue (1992): weights can be scaled instead of values.

- For the r-th replicate,

$$w_{hik}^{(r)} = \left\{ 1 - \left(\frac{m_h}{n_h - 1} \right)^{1/2} + \left(\frac{m_h}{n_h - 1} \right)^{1/2} \frac{n_h}{m_h} m_{hi}^{(*r)} \right\} w_{hik} \tag{4}$$

- $m_{hi}^{(*r)}$ = # times the i-th unit in stratum h is used in the r-th replicate
- Equivalent to RBS for functions of moments
- Applicable to $\hat{\theta}$ obtained from estimating equations

Bootstrap scheme options

Choice of m_h:

- $m_h \leq n_h - 1$ to ensure non-negative replicate weights
- $m_h = n_h - 1$: no need for internal scaling
- $m_h = n_h - 3$: matching third moments (Rao & Wu 1988)
- Simulation evidence (Kovar, Rao & Wu 1988): for $n_h = 5$, the choice $m_h = n_h - 1$ leads to more stable estimators with better coverage than $m_h = n_h - 3$

Choice of R:

- No theoretical foundations
- Popular choices: $R = 100$, 200 or 500
- $R \geq$ design degrees of freedom $= n - L$

Bootstrap without replacement (BWO)

BWO (Sitter 1992*a*) mimics sampling without replacement

1. Let $n_h^* = n_h - (1 - f_h)$, $k_h = \frac{N_h}{n_h}(1 - \frac{1-f_h}{n_h})$.
2. Create pseudopopulation: in stratum h, replicate $\{y_{hi}\}$ k_h times.
3. Take SRSWOR of n_h^* units from pseudopopulation stratum h, combine across h.
4. Compute $\hat{\theta}^{(r)}$.
5. Repeat Steps 3–4 for $r = 1, \ldots, R$.
6. Compute v_{BWO} using (2).
7. Randomize between bracketing integer values for non-integer n_h^*, k_h.

Extension to two-stage sample is available.

Mirror-match bootstrap (MMB)

MMB (Sitter 1992*b*) for sampling without replacement designs

1. Draw SRSWOR of $n_h^* < n_h$ PSUs from stratum h.
2. Repeat Step 1 $k_h = n_h(1 - f_h^*)/n_h^*(1 - f_h)$ times.
3. Repeat Steps 1–2 independently for each stratum to form the r-th replicate.
4. Compute $\hat{\theta}^{(r)}$.
5. Repeat Steps 1–4 for $r = 1, \ldots, R$.
6. Compute v_{MMB} using (2).

- $f_h = n_h/N_h$ is the original sampling fraction
- $f_h^* = n_h^*/n_h$ is the bootstrap sampling fraction
- Randomize if m_h/k_h is not integer
- Rescaling bootstrap: special case with $n_h^* = 1$

Mean bootstrap

Yung (1997), Yeo, Mantel & Liu (1999)

- Confidentiality protection: units with a weight of 0 belong to the same PSU, risk of identification.
- Replace the number of bootstrap draws $m_{hi}^{(*r)}$ by

$$\bar{m}_{hi}^{(*r)} = \frac{1}{K} \sum_{k=(r-1)K+1}^{rK} m_{hi}^{(*k)}$$

- Take K large enough so that $\Pr^*[\bar{m}_{hi}^{(*r)} = 0]$ is small.
- Proceed to compute the bootstrap weights (4).
- Compute v_{MBOOT} using (2) with scaling factor $A = K$.
- Number of resulting weight variables $= R/K$.

Warning: no formal theory have been developed so far.

Imputed data

Shao & Sitter (1996), Rao (1996); also JASA **91** (434)

Setup:

- $\mathbf{X}_R$ are the available responses
- $\mathbf{X}_M$ are the missing data
- A_R and A_M are indicators of complete/missing data
- Imputation procedure: $\eta_i = \mathcal{J}(\mathbf{X}_R; i), i \in A_M$
- $\mathbf{X}_I = \{x_i : i \in A_R\} \cup \{\eta_i : i \in A_M\}$ are imputed data

Imputed data

Shao & Sitter (1996): the bootstrap data set should be imputed in the same way as the original data set was!

1. Draw a bootstrap sample $(x^{(*r)}, a^{(*r)})$ of size $n_h - 1$ from $\mathbf{X}_I$ independently across strata h.
2. For resampled non-respondents $i \in A_M^{(*r)}$, apply the imputation procedure $\eta_i^{(*r)} = \mathcal{J}(\mathbf{X}_R^{(*r)}; i)$ to obtain re-imputed data set

$$X_I^{(*r)} = \{x_i^{(*r)} : i \in A_R^{(*r)}\} \cup \{\eta_i^{(*r)} : i \in A_M^{(*r)}\}$$

3. Compute $\hat{\theta}^{(r)}$.
4. Repeat Steps 1–3 for $r = 1, \ldots, R$.
5. Compute variance estimate using (2).

Balanced bootstraps

Nigam & Rao (1996)

- Special case: $m_h = n_h = n_0$ for all h
- Balance conditions:

$$\sum_r m_{hi}^{(*r)} = R, \quad n_0 \sum_r m_{hi}^{(*r)} m_{hj}^{(*r)} = R(n_0 \delta_{ij} + n_0 - 1),$$

$$n_0 \sum_r m_{hi}^{(*r)} m_{gk}^{(*r)} = R, g \neq h$$

- If $n_0 = 2m$ for some integer m, utilize Hadamard matrices and balanced incomplete block designs
- If $n_0 = p^k$ for prime p and integer k, utilize Hadamard matrices and Galois field theory

In general, the second order balance is very difficult to achieve.

Variance components bootstrap

Field & Welsh (2007): model-based survey inference

- Balanced random effects model:

$$Y_{ij} = \mu + \beta_i + \epsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m$$

- Goal: inference for σ^2_β, σ^2_ϵ
- Random effects bootstrap: sample β^*_i from $\{\hat\beta_i, i = 1, \ldots, n\}$, ϵ^*_{ij} from $\{\hat\epsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, m\}$
- Residual bootstrap:
  1. estimate $\hat\sigma^2_b$, $\hat\sigma^2_\epsilon$
  2. form $\hat C = I_n \otimes (\hat\sigma^2_\epsilon I_m + \hat\sigma^2_\beta J_m)$
  3. form whitened residuals $\mathbf{r} = \hat C^{-1/2}(\mathbf{y} - \hat\mu)$
  4. bootstrap $\mathbf{r}^*$ from $\mathbf{r}$
  5. form $\mathbf{y}^* = \hat\mu + \hat C^{1/2}\mathbf{r}^*$
- Cluster bootstrap: resample the whole cluster $\mathbf{Y}_i$
- All of the above are consistent when $n \to \infty$ with m fixed

Wild bootstrap

Cameron, Miller & Gelbach (2008)

- Regression model:

$$\mathbf{y}_i = \mathbf{X}_i \beta + \mathbf{u}_i$$

  where i enumerates clusters
- Goal: inference for $\hat{\beta}_{OLS}$
- Fit the model by OLS, obtain $\hat{\mathbf{u}}_i$
- Form the wild bootstrap samples by taking $\mathbf{u}_i^* = z_i^* \hat{\mathbf{u}}_i$, $\mathbf{Pr}[z_i^* = 1] = \mathbf{Pr}[z_i^* = -1] = 1/2$
- Applicable to both variance estimation and distribution estimation with bootstrap-t
- Simulation evidence: the wild cluster bootstrap outperforms other cluster bootstraps

Small area bootstrap

Lahiri (2003)

- For small area i, Y_i are observations, U_i are small area effects, X_i are regressors and Z_i are fixed constants
- Small area model:

$$Y_i|U_i \sim N(X_i\beta + Z_iU_i, R_i(\psi)), \quad U_i \sim N(0, G_i(\psi)) \quad (5)$$

- Quantity of interest:

$$\theta_i = l_i\beta + \lambda_i U_i$$

- BLUP/empirical Bayes predictor:

$$\hat{\theta}_i(Y_i;\hat{\psi}) = l_i\hat{\beta}(\hat{\psi}) + \lambda_i' G_i(\hat{\psi})(Z_i' G_i(\hat{\psi})Z_i + R_i(\hat{\psi}))^{-1}[Y_i - X_i\hat{\beta}(\hat{\psi})]$$

Small area bootstrap

- Inferential goal:

$$\mathrm{MSE}[\hat{\theta}] = \underbrace{g_1(\psi)}_{\text{due to } R_i \text{ and } G_i} + \underbrace{g_2(\psi)}_{\text{due to } \hat{\beta}} + \underbrace{\mathbb{E}[(\hat{\theta}_i(Y_i;\hat{\psi}) - \hat{\theta}_i(Y_i;\psi))^2]}_{\text{due to } \hat{\psi}}$$

- Parametric bootstrap from (5) with estimated parameters
- Bootstrap estimate:

$$\begin{aligned}\widehat{MSE}_{BS} &= g_1(\hat{\psi}) + g_2(\hat{\psi}) \\ &\quad - \mathbb{E}^*[g_1(\hat{\psi}^*) + g_2(\hat{\psi}^*) - g_1(\hat{\psi}) - g_2(\hat{\psi})] \\ &\quad + \mathbb{E}[(\hat{\theta}_i(Y_i^*;\hat{\psi}^*) - \hat{\theta}_i(Y_i^*;\hat{\psi}))^2]\end{aligned}$$

Big-O and small-o notation

Before we compare estimators, aside on notation.
For a (deterministic) sequence a_n, we shall write

- $a_n = O(n^\alpha)$ if $|a_n|/n^\alpha \leq M$ for sufficiently large n and M
- $a_n = o(n^\alpha)$ if $a_n/n^\alpha \to 0$

Examples:

- $\dfrac{\sin n}{n} = O(n^{-1}) = o(n^{-1/2})$
- $\log n = o(n^\alpha)$ for all $\alpha > 0$

Big-O and small-o notation

For a sequence of random variables V_n, we shall write

- $V_n = O_p(n^\alpha)$ if V_n/n^α is bounded in the limit in probability:

$$\forall \epsilon > 0 \, \exists M, n_0 < \infty : \mathbf{Pr}[|V_n|/n^\alpha > M] < \epsilon \text{ for } n \geq n_0$$

  for sufficiently large n and M
- $V_n = o_p(n^\alpha)$ if $V_n/n^\alpha \to 0$ in probability

Example: $\bar{X} \sim N(\mu, \sigma^2/n)$

- $\mathbb{V}[\bar{X}] = O(n^{-1})$
- $\bar{X} - \mu = O_p(n^{-1/2})$
- $v[\bar{X}] = s^2/n = O_p(n^{-1})$,
  $v[\bar{X}]/(\sigma^2/n) = 1 + O_p(n^{-1/2}) = 1 + o_p(1)$
- $t = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} = O_p(1)$

Comparisons of methods

Linear case: all estimators coincide!

$$v_L = v_{BRR} = v_J = v_{RBS} = v_{MMB} = v_{BWO} = v_{str}$$

The bootstrap methods are understood as the ideal/complete bootstrap. The actual applications may contain Monte Carlo bootstrap variability.

Linearization and the jackknife

Nonlinear case:

- The jackknife and linearization are consistent (Krewski & Rao 1981)
- The six jackknife variance estimators are equivalent up to $O_p(n^{-3})$ (Rao & Wu 1985)
- Relation to linearization:

$$v_J = v_L\left(1 + O_p(n^{-1})\right)$$

- If $n_h = 2$ for all h,

$$v_J = v_L\left(1 + O_p(n^{-2})\right)$$

Valliant (1996): v_J performs better than v_L in model-based approach to survey inference, ratio estimation, poststratification.

Linearization and BRR

In nonlinear case, BRR is consistent (Krewski & Rao 1981); relative accuracy (Rao & Wu 1985):

$$v_{BRR-H} = v_L\left(1 + O_p(n^{-1/2})\right)$$
$$v_{BRR-D} = v_L\left(1 + O_p(n^{-1})\right)$$
$$v_{BRR-S} = v_L\left(1 + O_p(n^{-1})\right)$$

Stability

Stability of estimators:

$$\text{rel.MSE}[v_m] = \frac{\mathbb{E}^{1/2}[(v_m - \sigma^2)^2]}{\mathbb{E}[(\hat{\theta} - \theta)^2]}$$

Simulation evidence (Krewski & Rao 1981, Rao & Wu 1988, Kovar, Rao & Wu 1988, Sitter 1992*a*):

1. Linearization (for smooth statistics)
2. The jackknife (for smooth statistics)
3. BRR
4. Bootstrap
   - MMB has a slight edge
   - RBS with $m_h = n_h - 3$ performs poorly

Confidence intervals

Simulation evidence for both one-sided and two-sided confidence intervals (Krewski & Rao 1981, Rao & Wu 1988, Kovar, Rao & Wu 1988, Sitter 1992*a*):

1. Bootstrap
   - MMB has a slight edge
   - RBS with $m_h = n_h - 3$ performs poorly
2. BRR
3. The jackknife (inconsistent for non-smooth statistics)
4. Linearization (inconsistent for non-smooth statistics)

Comparisons of methods

Shao (1996): "...the choice of the method may depend more on nonstatistical considerations, such as the feasibility of their implementation... Blindly applying the resampling methods may yield incorrect results"

Similar properties in pairs of methods:

- BRR is a special case of second-order balanced bootstrap
- Delete-1 jackknife and linearization are almost identical

Review questions

1. Which of the survey bootstrap methods has the most demanding memory requirements? The most demanding computational requirements?
2. How can stability of the bootstrap estimators improved?
3. Can the mean bootstrap be used for imputed data?
4. Give an example of an applied problem where the bootstrap will be preferred to linearization; linearization will be preferred to the bootstrap.

Notes

Software implementation

1. Cheat codes: bootstrap weights as BRR weights
2. Stata: `bsweights` and `bs4rw` packages
3. R: `survey` package, `svrepdesign` and `as.svrepdesign`

Cheat codes

Phillips (2004):

- The variance formula is (2) for both BRR and the bootstrap.
- Bootstrap weights as BRR weights!
- The mean bootstrap can be used with Fay's correction.

SUDAAN:

```
proc procname data=... design=BRR;
    weight = sampling weight;
    repwgt = bootstrap weights / adjfay = K
    ...
run;
```

Stata survey features

- `svy` suite of routines described in 170 pages manual
- Design flexibility: stratification, clustering, multiple stages of selection, probability weights, finite population corrections
- Estimation commands: totals, means, ratios, contingency tables with Rao-Scott corrections, GLM, microeconometrics, Cox regression
- Variance estimation: linearization, BRR, the jackknife
- Poststratification
- Post-estimation features: DEFF, MEFF, Wald tests

Complex survey bootstrap is implemented by `bsweights` (by Stas Kolenikov) in conjunction with `bs4rw` (by Jeff Pitblado of Stata Corp.).

`bsweights` syntax

`bsweights` *prefix*, `reps(#) n(#) [balanced replace calibrate(`*command* `@) verbose seed(#)]`

- `reps()` specifies the number of replications; required option.
- `n()` specifies the number of units to be resampled from each stratum; required option.
- `balanced` specifies balanced bootstrap.
- `calibrate` calls *command* substituting the name of the current replicate weight for @, and `verbose` shows the output of the calibrating command.
- `replace` allows overwriting the existing set of weights.
- `seed(#)` sets the pseudo-random number generator seed.

`bs4rw` syntax

`bs4rw` [*expression list*], `rweights(`*varlist*`)` [`vfactor(#)` `mse` *options*] : *command*

- `rweights(`*varlist*`)` specifies the bootstrap replicate variables; required option.
- `vfactor(#)` specifies the scaling factor A.
- `mse` requests to compute the MSE estimator; the default is to compute the variance estimator.
- *options*: output and reporting options.
- *command*: the estimation procedure to be bootstrapped; must contain the original sampling weights.

Bootstrap postestimation features (bias estimates and confidence intervals) are available with `estat bootstrap`.

Calibration

- Option `calibrate(`*call @*`)` allows to call an external program to perform additional adjustments on weights.
- The replication weight variables will be substituted for @ in the above call.
- Subpopulation estimation: set weights outside the subpopulation to zero.

Balancing conditions

First order balance can be achieved by `bsweights`:

- Each unit in stratum h is used the same number of times d_h.
- Total number of units used in all replications: $d_h n_h = m_h R$.
- Balancing condition: $m_h R$ is a multiple of n_h for all h; e.g., if n_h takes values 2, 3, 4 and 5, R must be a multiple of $3 \cdot 4 \cdot 5 = 60$.

Second order balance is difficult to satisfy for an arbitrary design.

Limitations

What `bsweights` **cannot** do:

- Design effect: a post-estimation feature, use Phillips (2004) trick
- Bootstrap t-percentiles of jackknife-after-bootstrap

$$\mathcal{D}[t] = \frac{\hat\theta - \theta}{\sqrt{v_J}} \quad \doteq \quad \mathcal{D}[t^*] = \frac{\hat\theta^* - \hat\theta}{\sqrt{v_J^*}}$$

- Finite population corrections
- Missing and imputed data: need a customized command to re-impute missing data and estimate the model
- Other survey bootstraps (MMB, BWO)

R implementation

R survey capabilities (Lumley forthcoming):

- Design flexibility: stratification, clustering, multiple stages of selection, probability weights, finite population corrections, two-phase designs
- Poststratification, raking and GREG calibration
- Estimation commands: totals, means, ratios, quantiles, contingency tables with Rao-Scott corrections, GLM, quasi-MLE, Cox regression
- Graphics for survey data
- Interface to `mi` package also written by the Thomas Lumley

R implementation

Replication variance estimation

- R, generating weights: `as.svrepdesign()`
  - The jackknife:
    `as.svrepdesign(design=...,type="JKn")`
  - BRR:
    `as.svrepdesign(design=...,type="BRR")`
  - Utilities to produce Hadamard matrices are included
  - RBS with $m_h = n_h - 1$:
    `as.svrepdesign(design=...,`
    `type="subbootstrap",replicates=...)`
  - Extract weights: `weights(design=...)`
- R, applying weights: `svrepdesign()`
  - `type=` as above
  - Separate data frames for variables, sampling weight and replication weights

Stata exercises

1. Run the bootstrap with and without balancing using several different seeds with the same number of replicates. Compare the results, including both the standard errors and bias estimates.
2. Modify the calibration program to calibrate the weights on gender and region.
3. (Lack of identification — trick question!) Provide the bootstrap analogue of

```
svy, subpop(region1): logistic highbp
female black orace
```

Run each of the examples to produce estimation output!

What I covered was. . .

1. Bootstrap for i.i.d. data
2. Variance estimation for complex surveys
3. Survey bootstraps
4. Software implementation
5. References

Question I don't know how to answer

- Can I use the survey bootstrap for multilevel models?
- How can I apply the bootstrap to longitudinal data?
- I am using [*the name of a complicated estimation procedure with several steps*]. Can the bootstrap be used to provide the standard errors?
- Do I really have to run $R = 500$ bootstrap replications for the imputed data bootstrap if I can get good results with $M = 5$ multiple imputations?

Overview questions

For a given situation, suggest the most appropriate variance estimation technique. Explain your choice.

1. You have collected some data on local businesses in a stratified one-stage equal probability sampling design with a handful of strata and several dozens observations in each stratum. You need to estimate several totals and proportions, and run a couple of regression analysis.
2. You are preparing a data set from a large scale survey for public release. The sampling design includes several hundreds PSUs arranged into strata, between 1 and 3 PSUs per stratum. You want the future users to be able to run any analysis they would need.
3. You are preparing an in-house report on income distribution and poverty for an existing large scale economics survey data with complex design. You have access to all the relevant design information, but you need to make sure that your report does not contain any information that could lead to confidentiality breaches.

Notes

Major references

- Shao (1996): a deep review of theory of resampling methods
- Rust & Rao (1996): a straightforward explanation of the methods with medical applications
- Rao & Wu (1988) and Rao, Wu & Yue (1992): major rescaling bootstrap papers
- Wolter (2007): in-depth advanced level coverage of all mathematical details

See list with links to full text at http://www.citeulike.org/user/ctacmo/tag/ce_jsm09_svy_bstrap

Complete list I

Andrews, D. F. (2007), 'Robust likelihood inference for public policy', *The Canadian Journal of Statistics* **35**(3), 341–350.

Andrews, D. W. K. (2000), 'Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space', *Econometrica* **68**(2), 399–405. doi:10.1111/1468-0262.00114.

Bickel, P. J. & Freedman, D. A. (1981), 'Some asymptotic theory for the bootstrap', *The Annals of Statistics* **9**(6), 1196–1217.

Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**, 279–292.

Cameron, C. A., Miller, D. & Gelbach, J. B. (2008), 'Bootstrap-based improvements for inference with clustered errors', *The Review of Economics and Statistics* **90**(3), 414–427.

Canty, A. J., Davison, A. C., Hinkley, D. V. & Ventura, V. (2006), 'Bootstrap diagnostics and remedies', *The Canadian Journal of Statistics/La revue canadienne de statistique* **34**(1), 5–27.

Complete list II

Davison, A. C., Hinkley, D. V. & Schechtman, E. (1986), 'Efficient bootstrap simulation', *Biometrika* **73**(3), 555–566.

Field, C. A. & Welsh, A. H. (2007), 'Bootstrapping clustered data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(3), 369–390.

Gleason, J. R. (1988), 'Algorithms for balanced bootstrap simulations', *The American Statistician* **42**(4), 263–266.

Graham, R. L., Hinkley, D. V., John, P. W. M. & Shi, S. (1990), 'Balanced design of bootstrap simulations', *Journal of the Royal Statistical Society* **52**(1), 185–202.

Gupta, V. K. & Nigam, A. K. (1987), 'Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum', *Biometrika* **74**(4), 735–742.

Gurney, M. & Jewett, R. S. (1975), 'Constructing orthogonal replications for variance estimation', *Journal of the American Statistical Association* **70**(352), 819–821.

Complete list III

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Hedayat, A., Sloane, N. J. A. & Stufken, J. (1999), *Orthogonal Arrays: Theory and Applications*, Springer Series in Statistics, Springer-Verlag, New York.

Kish, L. & Frankel, M. R. (1974), 'Inference from complex samples', *Journal of the Royal Statistical Society, Series B* **36**, 1–37.

Kovar, J. G., Rao, J. N. K. & Wu, C. F. J. (1988), 'Bootstrap and other methods to measure errors in survey estimates', *Canadian Journal of Statistics* **16**, 25–45.

Krewski, D. & Rao, J. N. K. (1981), 'Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods', *The Annals of Statistics* **9**(5), 1010–1019.

Lahiri, P. (2003), 'On the impact of bootstrap in survey sampling and small-area estimation', *Statistical Science* **18**, 199–210.

Lumley, T. (forthcoming), *Complex Surveys: A guide to analysis using R*, Wiley, New York.

Complete list IV

McCarthy, P. J. (1969), 'Pseudo-replication: Half samples', *Review of the International Statistical Institute* **37**(3), 239–264.

Nigam, A. K. & Rao, J. N. K. (1996), 'On balanced bootstrap for stratified multistage samples', *Statistica Sinica* **6**(1), 199–214.

Phillips, O. (2004), Using bootstrap weights with WesVar and SUDAAN, Technical Report 2, Statistics Canada.

Rao, J. N. K. (1996), 'On variance estimation with imputed survey data', *Journal of the American Statistical Association* **91**(434), 499–506.

Rao, J. N. K. & Wu, C. F. J. (1985), 'Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics', *Journal of the American Statistical Association* **80**(391), 620–630.

Rao, J. N. K. & Wu, C. F. J. (1988), 'Resampling inference with complex survey data', *Journal of the American Statistical Association* **83**(401), 231–241.

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992), 'Some recent work on resampling methods for complex surveys', *Survey Methodology* **18**(2), 209–217.

Complete list V

Rust, K. F. & Rao, J. N. (1996), 'Variance estimation for complex surveys using replication techniques', *Statistical Methods in Medical Research* **5**(3), 283–310.

Shao, J. (1996), 'Resampling methods in sample surveys (with discussion)', *Statistics* **27**, 203–254.

Shao, J. & Sitter, R. R. (1996), 'Bootstrap for imputed survey data', *Journal of the American Statistical Association* **91**(435), 1278–1288.

Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, Springer, New York.

Sitter, R. R. (1992*a*), 'Comparing three bootstrap methods for survey data', *The Canadian Journal of Statistics* **20**(2), 135–154.

Sitter, R. R. (1992*b*), 'A resampling procedure for complex survey data', *Journal of the American Statistical Association* **87**(419), 755–765.

Sitter, R. R. (1993), 'Balanced repeated replications based on orthogonal multi-arrays', *Biometrika* **80**(1), 211–221.

Valliant, R. (1996), 'Discussion of "Resampling Methods in Sample Surveys" by J. Shao', *Statistics* **27**, 247–251.

Complete list VI

Wolter, K. M. (2007), *Introduction to Variance Estimation*, 2nd edn, Springer, New York.

Wu, C. F. J. (1986), 'Jackknife, bootstrap and other resampling methods in regression analysis', *The Annals of Statistics* **14**(4), 1261–1295.

Wu, C. F. J. (1991), 'Balanced repeated replications based on mixed orthogonal arrays', *Biometrika* **78**(1), 181–188.

Yeo, D., Mantel, H. & Liu, T.-P. (1999), Bootstrap variance estimation for the National Population Health Survey, *in* 'Proceedings of Survey Research Methods Section', The American Statistical Association, pp. 778–785.

Yung, W. (1997), Variance estimation for public use files under confidentiality constraints, *in* 'Proceedings of Statistics Canada Symposium', Statistics Canada, pp. 434–439.

Notation I

The generic datum x_{hij} denotes the measurement on variable y taken on the j-th observation in the i-th PSU in stratum h.

h	stratum index; $h = 1, \ldots, L$
i	PSU index within strata; $i = 1, \ldots, n_h$
j	observation index within PSU
L	number of strata
m_h	bootstrap sample size; the number of PSUs resampled from stratum h
$m_{hi}^{(*r)}$	bootstrap frequency; the number of times unit h, i is sampled in the r-th replicate
n	total sample size; the total number of PSUs in the sample: $n = \sum_{h=1}^{L} n_h$
N	population size; the total number of PSUs in the population: $N = \sum_{h=1}^{L} N_h$
n_h	sample size in stratum h; the number of PSUs taken from stratum h
N_h	population size; the number of PSUs in stratum h

Notation II

R	the number of replicates; the number of replicate weights in the mean bootstrap
$T[x]$	population total: $T[x] = \sum_h \sum_i \sum_j x_{hij}$
$t[x]$	estimate of the population total $T[x]$
$v[\hat{\theta}]$	estimator of variance $\mathbb{V}[\hat{\theta}]$
$v_m[\hat{\theta}]$	estimator of variance $\mathbb{V}[\hat{\theta}]$ obtained by method m; the methods include linearization L, the jackknife J, BRR, or bootstrap schemes RBS, MMB, BWO and BWR
$\mathbb{V}[\hat{\theta}]$	(design) variance of the estimate $\hat{\theta}$ with respect to the sampling distribution
W_h	fraction of stratum h in population: $W_h = N_h/N$
w_{hij}	sampling weight of unit h, i, j
$w_{hij}^{(r)}$	replicate weight of unit h, i, j in the r-th replicate
θ	population parameter, such as total, mean, ratio, regression coefficient
$\hat{\theta}$	parameter estimate obtained from survey data
$\hat{\theta}^{(r)}$	parameter estimate obtained in the r-th replicate

Notation III

BRR	balanced repeated replication	39
BWO	bootstrap without replacement	50
MSE	mean-squared error	
NBS	naïve bootstrap	48
MMB	mirror-match bootstrap	53
PSU	primary sampling unit	26
RBS	rescaling bootstrap	49
SRSWOR	simple random sample without replacement	

Survey bootstrap theory: SRSWR I

- Design: SRSWR sampling n out of N
- Population parameters:

$$\bar{x}_U = \frac{1}{N}\sum_{i\in\mathcal{U}} x_i, \quad \sigma^2 = \frac{1}{N}\sum_{i\in\mathcal{U}}(x_i - \bar{x})^2$$

- Sample statistics:

$$\bar{x} = \frac{1}{n}\sum_{i\in\mathcal{S}} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(x_i - \bar{x})^2$$

- Bootstrap sample: SRSWR n^* out of n units, x_i^* from $\{x_1, \ldots, x_n\}$

Survey bootstrap theory: SRSWR II

- Ideal bootstrap variance:

$$\mathbb{V}^*[x_1^*] = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{n-1}{n}s^2$$

$$\mathbb{V}^*[\bar{x}] = \frac{1}{n^*}\mathbb{V}^*[x_1^*] = \frac{n-1}{nn^*}s^2 \neq v[\bar{x}] = \frac{s^2}{n}$$

- Unbiased: only if $n^* = n$

Survey bootstrap theory: SRSWOR I

- Design: SRSWOR sampling n out of N
- Population parameters:

$$\bar{x}_U = \frac{1}{N}\sum_{i\in\mathcal{U}} x_i, \quad S^2 = \frac{1}{N-1}\sum_{i\in\mathcal{U}}(x_i - \bar{x})^2$$

- Sample statistics:

$$\bar{x} = \frac{1}{n}\sum_{i\in\mathcal{S}} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(x_i - \bar{x})^2$$

- Bootstrap sample: SRSWR n^* out of n units, x_i^* from $\{x_1, \ldots, x_n\}$

Survey bootstrap theory: SRSWOR II

- Ideal bootstrap variance:

$$\mathbb{V}^*[x_1^*] = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{n-1}{n}s^2$$

$$\mathbb{V}^*[\bar{x}] = \frac{1}{n^*}\mathbb{V}^*[x_1^*] = \frac{n-1}{nn^*}s^2 \neq v[\bar{x}] = (1-f)\frac{s^2}{n}$$

- Solutions:
  - scale the variance by $(1-f)n^*/(n-1)$
  - use internal scaling (Rao & Wu 1988)
  - use special algorithms (BWR, BWO, MMB)

Jump back to the middle of the notes

The end

- Questions? Clarifications? Additional help? Email me: kolenikovs@missouri.edu
- Please fill ASA course evaluation forms.
- If you liked this course, invite me to give it at your organization.

THANKS!