Introduction to Bootstrapping Simulation in SAS

Yubo Gao, PhD
Biostatistician
Orthopaedic and Rehabilitation
August 21, 2014

Simulation

- A process where we estimate parameter through resampling our original sample
- Resampling methods:

  (1) Bootstrapping

  (2) Jackknifing, and

  (3) Others

Bootstrapping

- A non-parametric method-no distribution assumption
- A simple random sampling(SRS) with replacement-also called Unrestricted

  Random Sampling (URS) in SAS
- Used to estimate standard error, confidence interval, mean, median, and others
- For bootstrapping theory, see Efron's works
- Here, we focus on its implementation in SAS through examples

An Old Bootstrapping Way

```
%macro old_boot ( input=, reps= ); /*to calculate 95% CI kurtosis*/

 %do i = 1 %to &reps ;
          data generated;
            do i=1 to nobs;
                rec = ceil(nobs * ranuni(0));
                set &input nobs=nobs point=rec;
                output;
            end;
           run;

          proc univariate data=generated;
              var x;
              output out=outx kurtosis=curt;
          run;
```

An Old Bootstrapping Way-cont.

```
        %if &i = 1 %then %do;
            data outall;
                set outx;
            run;
           %end;
        %else %do;
            proc append base=outall data=outx;
            run;
        %end;
%end; /* i=1 to &reps loop */
```

An Old Bootstrapping Way-cont.

```
        proc univariate data=outall;
                var curt;
                output out=final pctlpts=2.5, 97.5 pctlpre=ci;
        run;

        proc print;
        run;

%mend;

%old_boot(input=YourData, reps=1000);
```

Disadvantages

- Seed generated internally by SAS, thus no way to reproduce the results
- The POINT= method picks record set in a non-sequential manner, which may be much slower than the usual SAS sequential reading of data sets
- And the NOBS= option does not always work

A Better Bootstrapping Way—using **proc surveyselect**

- Note that it can generate many kinds of samples.
- Here we just focus on simple random samples, or called Unrestricted Random Sampling (URS) in SAS

A Better Bootstrapping Way—using **proc surveyselect-cont.**

```
proc surveyselect data=YourData out=outboot
        seed=30459584
        method=urs          /* specify the type of random sampling */
        samprate=1          /* get a sample of the same size as
                               our original data set */
        outhits             /*give the times a record chosen*/
        rep=1000;           /* specify the number of bootstrap samples
                               that we want */
run;

proc univariate data=outboot;
        var x;
        by Replicate;
        output out=outall kurtosis=curt; /* ODS does the same  thing*/
run;
```

A Better Bootstrapping Way—using **proc surveyselect-cont.**

/*This is the Percentile Interval. Other Intervals exist that need little more coding efforts,such as the Bootstrap-T, the Bias-corrected and Accelerated (BCa) intervals.*/

```
proc univariate data=outall;
        var curt;
        output out=final pctlpts=2.5, 97.5 pctlpre=ci;
run;

proc print;
run;
```

An Example

```
OPTIONS nocenter MPRINT SYMBOLGEN
formdlim=' ';

data one;
        input sales @@;
datalines;
16065 23868 26881 92114 93132 93136 94075
94137 94306 95031 95217 98530 99037 99136
99818
;
run;
```

An Example-cont.

```
proc surveyselect data=one out=outboot
    seed=30459584
    method=urs  /* type of random sampling */
    samprate=1  /* get a sample of the same size as
                   our original data set */
    outhits        /*give the times a record chosen*/
    rep=100;      /*  number of bootstrap samples
                   that we want */
run;
```

An Example-cont.

```
proc univariate data=outboot noprint;
        var sales;
        by Replicate;
        output out=outall kurtosis=curt;
run;

proc univariate data=outall;
        var curt;
        output out=final pctlpts=2.5, 97.5 pctlpre=ci;
run;

proc print;
run;
```

An Application

- A Risk Calculator for Short Term Morbidity and Mortality Following Hip Fracture Surgery.

  Journal of Orthopaedic Trauma. 18 July, 2013

  Andrew J. Pugely, Christopher T. Martin, Yubo Gao, Lawrence Marsh, John J. Callaghan
- Also presented at SAS Global Forum 2014 in DC, March 2014

Introduction

- Hip fractures are a common source of morbidity and mortality amongst the elderly
- Few studies have presented a validated method for stratifying patient risk.
- The purpose was to develop a simple risk score calculator predictive of 30-day morbidity after hip fracture.

Data Aquisition

- Data are from ACS NSQIP-(American College of Surgeons National Surgical Quality Improvement Program with all cases of hip fracture between 2005 and 2010,
- Based on primary Current Procedural Terminology (CPT) codes.
- Total 4,331 patients selected.

Variables

- Demographic data : age, sex, race
- Pre-operative health variables
- Pre-operative comorbidities
- Pre-operative laboratory values
- Operative variables
- Outcomes-30-day complications after hip fracture
  1) mortality- death within 30 days
  2) complications-infectious, pulmonary, hematologic , cardiac, renal, neurologic, hardware failure, return to the OR within 30 days.
  3) complication was defined by the presence of one or more of the above positive outcomes

Statistical Methods

- Univariate analysis
- Multi-variate logistic regression
- Bootstrap simulation
- Divide whole dataset into 80% development dataset (80%) and validation dataset (20%)

Statistical Methods-cont.

- Once validated, all pre-operative variables identified by the multivariate logistic regression analysis were introduced into a risk score stratification model.

- A 200-cycle bootstrapped simulation sample was used to generate beta coefficients of each risk factor included in the logistic regression model for the development data set.

Statistical Methods-cont.

```
proc surveyselect data=data80 out=outboot
        seed=30459584
        method=urs
        samprate=1
        outhits
        rep=200;
run;
```

Statistical Methods-cont.

```
ODS OUTPUT ParameterEstimates=allbeta LackFitChiSq= LackFit;
    proc logistic data=outboot descending;
        by Replicate;
        class  sex(param=ref ref='female')
               fnstatus10(param=ref ref='Independent')
               dyspnea1(param=ref ref='No')
               hxchf(param=ref ref='No')    /*heart failure*/
               nd(param=ref ref='No')        /*neurological disease*/
               wbc(param=ref ref='2');
        model majorcomplication=sex fnstatus10 dyspnea1 hxchf nd
                                    wbc/lackfit risklimits;

    run;
    quit;
ODS OUTPUT CLOSE;
```

Statistical Methods-cont.

```
/*Some data manipulations*/
data beta1;
        set allbeta;
        variable_name=compress(variable||ClassVal0,'');
run;

/*Change format so that each row is one record for each sample*/
proc transpose data=beta1  out=beta3;
        by Replicate;
        id Variable_name;
        var estimate;
run;
```

Statistical Methods-cont.

```
proc contents data=beta3 out=all2;
run;

/*Pull out variable names into macro variable NameID to be used in proc means*/
proc sql noprint;
        select name into:NameID separated by ' ' from all2;
run;

%put &NameID;    /*check variables*/

title 'Beta Medians distributions of each variable';
proc means data=beta3 n mean median std range maxdec=2;
        var &NameID;
run;
```

Statistical Methods-cont.

- Next after validating the model using data20, then assign each score based on the beta medians, and determine the risk score for each patient by summing the individual risk factor points.

- The risk scores were further stratified as follows: (1) low, 0 to 10; (2) low-moderate, 11 to 15; (3) moderate-high, 16 to 20; and (4) high, 21+.

- Now it is ready for patient to use at http://www.uiortho.com/races/index.php/test

Conclusions

- Established a unique model for quantifying the impact of patient risk factors on outcomes after the surgical treatment of hip fractures.

- The risk score calculators were internally validated, and would be useful in informing patient discussions about operative risk, for identifying patients in need of medical optimization pre-operatively, and for comparing risk adjusted outcomes between surgeons or institutions

More about Bootstrapping

- This is just a very short introduction.
- It can be applied in many situations.
- For example, comparing the effectiveness between two treatments, Kappa coefficient estimation, and *et al*.
- Especially for small sample.

Any suggestions/comments.

Thank you!

Yubo Gao, PhD
Biostatistician
Orthopaedics & Rehabilitation
01066 JPP
yubo-gao@uiowa.edu
356-1674