Resampling and the Bootstrap

Axel Benner
Biostatistics, German Cancer Research Center
INF 280, D-69120 Heidelberg
benner@dkfz.de

Topics

- Estimation and Statistical Testing
  - Simulation
  - Bootstrap
  - Jackknife
  - Permutation

- Prediction
  - Jackknife
  - Cross-validation
  - Bootstrap

Resampling

- Approximations obtained by random sampling or simulation are called **Monte Carlo** estimates

  Assume random variable Y has a certain distribution
  $\rightarrow$ Use simulation or analytic derivations to study how an estimator, computed from samples from this distribution, behaves.

  e.g. Y has lognormal distribution $\Rightarrow$ var(median) = ?

  1. analytical solution?
  2. simulate 500 samples of size n from the lognormal distribution, compute the sample median for each sample, and then compute the sample variance of the 500 sample medians.

  Problem: need knowledge of the population distribution function

Example of 100 random deviates

```
?rlnorm
n <- 100
set.seed(12345)
y <- matrix(rlnorm(500*n,meanlog=0,sdlog=1),nrow=n,ncol=500)
summary(apply(log(y),2,mean))
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.307200 -0.060810 -0.001447  0.003774  0.076070  0.268100
```

```
summary(apply(log(y),2,sd))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7787  0.9433  0.9941  0.9945  1.0440  1.2190
```

```
ym <- apply(y, 2, median)
print(var(ym))
```

```
[1] 0.01567178
```

Baron Münchhausen

- He found himself at the bottom of a lake (swamp)
- ... and pulled himself up by his bootstraps!

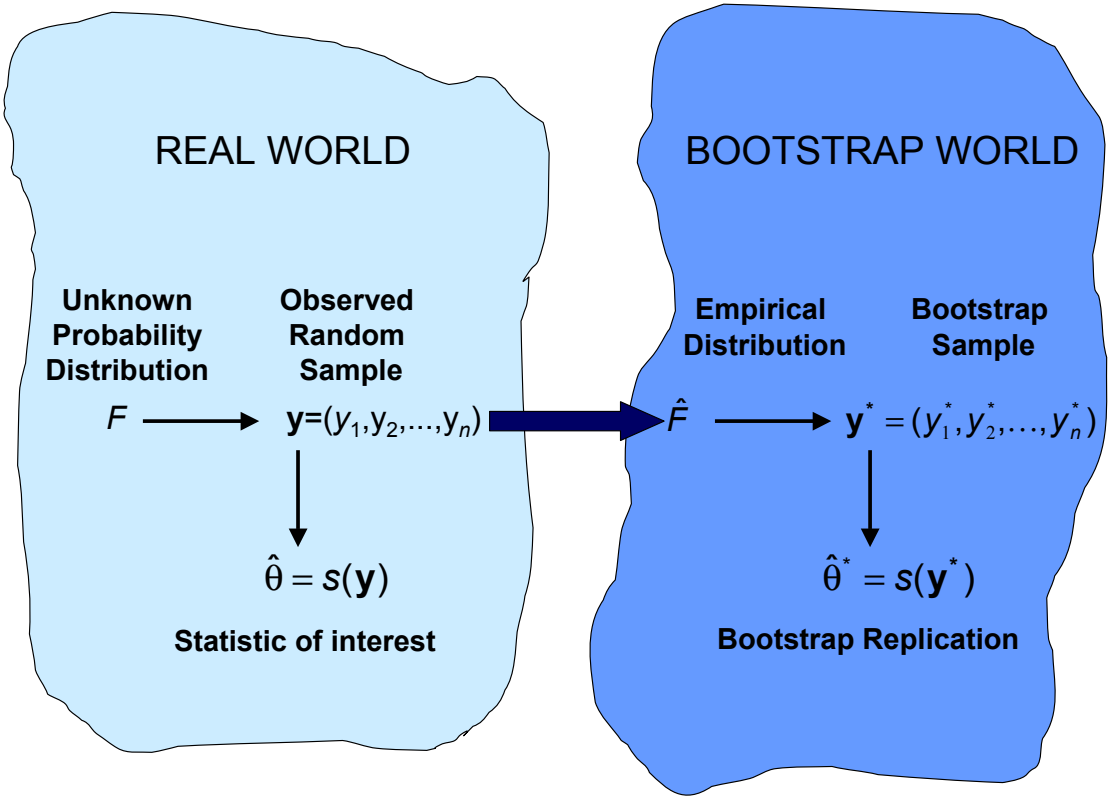
- No fraud!!

The Bootstrap

Efron's bootstrap is a general purpose technique for obtaining estimates of properties of statistical estimators without making assumptions about the distribution of the data.

Often used to find

1. standard errors of estimates
2. confidence intervals for unknown parameters
3. p values for test statistics under a null hypothesis


REAL WORLD
Unknown Probability Distribution
Observed Random Sample
F
y=(y1,y2,...,yn)
θ̂ = s(y)
Statistic of interest
BOOTSTRAP WORLD
Empirical Distribution
Bootstrap Sample
F̂
y* = (y1*, y2*, ..., yn*)
θ̂* = s(y*)
Bootstrap Replication

The Bootstrap

Suppose Y has a cumulative distribution function (cdf) $F(y) = P(Y \leq y)$

We have a sample of size n from $F(y)$, $Y_1, Y_2, \ldots, Y_n$

Steps:

1. Repeatedly simulate sample of size n from F
2. Compute statistic of interest
3. Study behavior of statistic over B repetitions

- Without knowledge of F we use the empirical cdf $F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$ as an estimate of F.

- Pretend that $F_n(y)$ is the original distribution $F(y)$.

- Sampling from $F_n(y)$ is equivalent to sampling with replacement from originally observed $Y_1, ..., Y_n$

- For large n the expected fraction of original data points that are selected for each bootstrap sample is 0.632

$$\begin{aligned} P(\text{obs. } i \in \text{ bootstrap sample } b) &= 1-(1-\frac{1}{n})^n \\ &\approx 1-e^{-1} \\ &= 0.632 \end{aligned}$$

Note: $1-\frac{1}{n}$ is probability for not being selected at a specific drawing; with n drawings we get that $(1-\frac{1}{n})^n$ is probability of not being selected at least once.

- From bootstrap sampling we can estimate any aspect of the distribution of $s(y)$ [which is any quantity computed from the data y], for example its standard error

$$\widehat{se}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2 \right\}^{1/2}$$

where $\hat{\theta}^*(b) = s(y^{*b})$ is the bootstrap replication of $s(y)$ and $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$.

The Jackknife ("a large strong pocketknife", Quenouille, 1949)

- We have a sample $y = (y_1, ..., y_n)$ and estimator $\hat{\theta} = s(y)$.
- Target: Estimate the bias and standard error of $\hat{\theta}$.
- The leave-one-out observation samples

$$y_{(i)} = (y_1, ..., y_{i-1}, \quad y_{i+1}, ..., y_n)$$

  for $i = 1, ..., n$ are called **jackknife samples**.

- Jackknife estimators are $\hat{\theta}_{(i)} = s(y_{(i)})$.

- The jackknife estimate of bias is

$$\widehat{bias}_J = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

  where $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$

- The jackknife estimate of the standard error is

$$\hat{se}_J = \sqrt{\frac{n-1}{n}\sum(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}$$

- The Jackknife often provides a simple and good approximation to the bootstrap (see below) for estimation of standard error and bias.

- But can fail if $\hat{\theta}$ is not "smooth" (i.e. differentiable)!

Example Accuracy of the median

```
y <- scan()
10 27 31 40 46 50 52 104 146

median(y)

[1] 46
```

Note that the median is not a differentiable (i.e. smooth) function of y.

Increasing the 4th smallest value $y = 40$ does not change the median until y exceeds 46, after that, the median is equal to y , until y exceeds 50!

Example Accuracy of the median (cont)

The jackknife values of the median are

```
n <- length(y)
ymat <- matrix(y, n-1, n)
y.j <- rev(apply(ymat, 2, median))
print(y.j)
```

```
[1] 48 48 48 48 45 43 43 43 43
```

Note that there are only 3 distinct values.

Example Accuracy of the median (cont)

Now compare it with the bootstrap estimate using 200 bootstrap samples

```
se.j <- sqrt( (n-1)/n * sum( (y.j -mean(y.j))^2 ))
set.seed(54321)
ymat <- matrix(sample(y, size=n*200, replace=T), n, 200)
y.b <-  apply(ymat, 2, median)
se.b <- sd(y.b)
print(c(se.j, se.b))
```

```
[1]  6.681465 11.313352
```

- Fix the inconsistency by using *delete-d jackknife*
- Use jackknife to measure the uncertainty of the bootstrap estimate of the standard error of a statistic $s(y)$: *jackknife-after-bootstrap* $\longrightarrow$ cp. function `jack.after.boot` from R package `boot`.

Sensitivity analysis (Jackknife after Bootstrap)

How different would the results have been if an observation y_j has been absent from the original data?

Measure effect of y_j on calculations by comparing full simulation with the subset of statistics $t_1^*, ..., t_B^*$ obtained from bootstrap samples without y_j.

Using frequencies f_{bj}^* counting the number of times y_j appears in the bth simulation we restrict to replicates with $f_{bj}^* = 0$.

$\Rightarrow$ Measure effect of y_j on the bias by scaled difference

$$n(bias_{-j} - bias) = \left\{ \frac{1}{B_{-j}} \sum_{b:f_{bj}^*=0} (t_b^* - t_{-j}) - \frac{1}{B} \sum (t_b^* - t) \right\}$$

t_{-j} is the value of t when y_j is excluded from the original data.

Hypothesis testing

- Null hypothesis: absence of some effect
- Hypothesis test
  - within the context of a statistical model
  - without a model
    (1) non-parametric
    (2) permutation
- Many hypotheses
  - Testing too many hypotheses is related to fitting too many predictors in a regression model
  - Point estimates are badly biased when the quantity to be estimated was determined by "data dredging" (cp. shrinkage).
    This is especially true when one wants to publish an effect estimate corresponding to the hypothesis yielding the smallest p-value

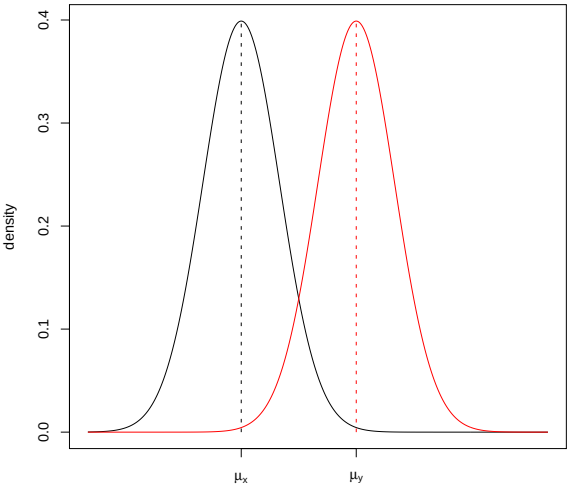
Classical inference statistic: parametric statistical tests

- statistical hypothesis concerning the distribution of features of interest in a population (or its parameters)
- checking the hypothesis by random sampling from the population
- Assumption:
  every element of the population has the same chance to be included into the sample
- a statistical test is based on a test statistic T which measures the discrepancy between the data and the null hypothesis

Permutation tests

- Also called *randomization tests, rerandomization tests, exact tests.*
- Introduced in the 1930s.
- Usually require only a few weak assumptions.

Example: t-test for independent samples



- Comparison of expectations in two normal distributed populations
- Assumptions:
  two normal distributed populations with means μ_x and μ_y and identical variances $\sigma_x^2 = \sigma_y^2$

Null hypothesis: $H_0 : \mu_x = \mu_y$
Alternative hypothesis (two-sided): $H_A : \mu_x \neq \mu_y$

- random samples $X = \{x_1 \ldots, x_{n_x}\}$ and $Y = \{y_1, \ldots, y_{n_y}\}$
- test statistic

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{(n_x-1)+(n_y-1)}}} \cdot \sqrt{\frac{n_x \cdot n_y}{n_x + n_y}}$$

  ($\overline{X}$, $\overline{Y}$ and S_x^2, S_y^2 are sample means and sample variances, n_x, n_y sample sizes)

- under H_0 holds $T \sim t_{n_x+n_y-2}$ (t-distribution with $n_x + n_y - 2$ d.f.)
- compute the p-value for the observed value t of test statistic T

$$p = 1-P(|T| \leq |t||H_0) = 2[1-P(T \leq |t||H_0)] = 2[1-F_{t,n_x+n_y-2}(|t|)]$$

- Decision rule: reject H_0 if p-value $\leq \alpha$

The p-value

- The p-value is the chance of obtaining a test statistic as or more extreme (as far away from what we expected or even farther in the direction of the alternative) than the one we got, assuming the null hypothesis is true.
- This chance is called the observed significance level, or p-value.
- A test statistic with a p-value less than some prespecified false positive level (or size) α is said to be 'statistically significant' at that level.
- The p-value represents the probability that we would observe a difference as large as we saw (or larger) if there were really nothing happening other than chance variability.
- The *significance level* (or size) α of a test is the probability of making a Type I error; that is, α is the probability of deciding erroneously on the alternative when, in fact, the hypothesis is true.

The power of a test

- Two types of error

	Decision	
The Facts	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$
$\mu_x = \mu_y$		Type I error α
$\mu_x \neq \mu_y$	Type II error β	

- The *power* $1 - \beta$ of a test is 1 minus the probability of making a type II error; that is, $1 - \beta$ is the probability of deciding on the alternative when the alternative is the correct choice.
- The ideal statistical test would have a significance level α of zero and a power of 1, but this ideal can not be realized.
- In practice, we will fix a significance level $\alpha > 0$ (usually this will be 0.05), and choose a statistic that maximizes or comes closest to maximizing the power $1 - \beta$.

Five steps to a Permutation Test

1. Analyze the problem
   - What is the hypothesis? What are the alternatives?
   - What distribution is the data drawn from?
   - What losses are associated with bad decisions?
2. Choose a test statistic which will distinguish the hypothesis from the alternative.
3. Compute the test statistic for the original labelling of the observations.
4. Compute the test statistic for all possible permutations (rearrangements) of the labels of the observations.
5. Make a decision

   Reject the hypothesis and accept the alternative if the value of the test statistic for the original labelling (original data) is an extreme value in the permutation distribution of the statistic. Otherwise, accept the hypothesis and reject the alternative.

Example t test vs. permutation test

- data

X			Y		
A	B	C	D	E	F
121	118	110	34	12	22
$\overline{x}_n = 116.33$			$\overline{y}_n = 22.67$		

  t test statistic: $t = 13.0875$, two-sided p-value: $p = 0.0002$

- after one permutation:

X			Y		
A	B	D	C	E	F
121	118	34	110	12	22
$\overline{x}_n = 91$			$\overline{y}_n = 48$		

- how many permutations exist?

$$C_3^6 = \binom{6}{3} = \frac{6!}{3! \cdot 3!} = \frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3} = 20$$

permutation	X	Y	$\overline{x}_n$	$\overline{y}_n$	$\overline{x}_n - \overline{y}_n$	t
1	ABC	DEF	116.33	22.67	93.67	13.087
2	ABD	CEF	91.00	48.00	43.00	1.019
3	ABE	CDF	87.00	52.00	35.00	0.795
4	ABF	CDE	83.67	55.33	28.33	0.627
5	ACD	BEF	88.33	50.67	37.67	0.866
6	ACE	BDF	84.33	54.67	29.67	0.659
7	ACF	BDE	81.00	58.00	23.00	0.500
8	ADE	BCF	59.00	80.00	-21.00	-0.455
9	ADF	BCE	55.67	83.33	-27.67	-0.611
10	AEF	BCD	51.67	87.33	-35.67	-0.813
11	BCD	AEF	87.33	51.67	35.67	0.813
12	BCE	ADF	83.33	55.67	27.67	0.611
13	BCF	ADE	80.00	59.00	21.00	0.455
14	BDE	ACF	58.00	81.00	-23.00	-0.500
15	BDF	ACE	54.67	84.33	-29.67	-0.659
16	BEF	ACD	50.67	88.33	-37.67	-0.866
17	CDE	ABF	55.33	83.67	-28.33	-0.627
18	CDF	ABE	52.00	87.00	-35.00	-0.795
19	CEF	ABD	48.00	91.00	-43.00	-1.019
20	DEF	ABC	22.67	116.33	-93.67	-13.087

- Test decision: In two of 20 cases overall the absolute value of the test statistic t is greater than or equal to the absolute value of $t = 13.087$ we obtained for the original labelling.

  Therefore we obtain the exact p value $p = 2/20 = 0.1$.

- Note: 0.1 is the smallest p value you can get for comparing two groups of size 3.
- Note: If both groups have equal size only half of permutations is really needed (symmetry)
- Note: The number of permutations for comparing two groups of size m and $n - m$ is

$$C_m^n = \binom{n}{m} = \frac{n!}{m! \cdot (n-m)!}$$

e.g. for $n = 52$ and $m = 18$

$$C_{18}^{52} = \binom{52}{18} = \frac{52!}{18! \cdot 34!} = 4.27 \times 10^{13}$$

- It may be necessary to use Monte Carlo sampling to approximate the permutation test

Microarray Data

Estimate the joint distribution of the test statistics $T_1, \ldots, T_G$ under the complete null hypothesis H_0^C by permuting the columns of the $(G \times n)$ gene expression data matrix $\mathbf{X}$.

Permutation algorithm for non-adjusted p-values

- For the b-th permutation, $b = 1, \ldots, B$
  1. Permute the n columns of the data matrix $\mathbf{X}$.
  2. Compute test statistics $t_{1,b}, \ldots, t_{G,b}$ for each hypothesis.
- The permutation distribution of the test statistic T_g for hypothesis H_g, $g = 1, \ldots, G$, is given by the empirical distribution of $t_{g,1}, \ldots, t_{g,B}$. For two-sided alternative hypotheses, the permutation p-value for hypothesis H_g is

$$p_g^* = \frac{1}{B} \sum_{b=1}^{B} I(|t_{g,b}| \geq |t_g|)$$

where $I(.)$ is the indicator function, equaling 1 if the condition in parenthesis is true, and 0 otherwise.

Permutation algorithm of Westfall & Young (maxT)

- Order observed test statistics: $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_G}|$.
- For the b-th permutation of the data ($b = 1, \ldots, B$):
  - divide the data into its artificial control and treatment group
  - compute test statistics $t_{1,b}, \ldots, t_{G,b}$
  - compute successive maxima of the test statistics

$$\begin{aligned} u_{G,b} &= |t_{r_G,b}| \\ u_{g,b} &= \max\{u_{g+1,b}, |t_{r_g,b}|\} \text{ for } g = G-1, \ldots, 1 \end{aligned}$$

- compute adjusted p-values:

$$\tilde{p}^*_{r_g} = \frac{1}{B} \sum_{b=1}^{B} I(u_{g,b} \geq |t_{r_g}|)$$

Permutation algorithm of Westfall & Young – Example

gene	|t|	
1	0.1	t_{r_G}
4	0.2	$t_{r_{G-1}}$
5	2.8	:
2	3.4	t_{r_2}
3	7.1	t_{r_1}

sort observed values

gene	$\|t_b\|$	u_b	$I(u_b>\|t\|)$
1	1.3	1.3	1
4	0.8	1.3	1
5	3.0	3.0	1
2	2.1	3.0	0
3	1.8	3.0	0

B=1000 permutations

$\sum$	$\tilde{p} = \sum / B$
935	0.935
876	0.876
138	0.138
145	0.145
48	**0.048**

adjusted p-values

O. Hartmann - NGFN Symposium, 19.11.2002 Berlin

Nonparametric Bootstrap Tests

- permutation tests are special nonparametric resampling tests, in which resampling is done without replacement
- the special nature of significance tests requires that probability calculations be done under a null hypothesis model, that means we must resample from a distribution $\hat{F}_0$, say, which satisfies the relevant null hypothesis H_0
- the basic bootstrap test will be to compute the p-values as

$$p_{boot} = P^*(T^* \geq t | \hat{F}_0)$$

approximated by

$$p = \frac{1}{B} \sum_{b=1}^{B} I(t_b^* \geq t)$$

using the results $t_1^*, t_2^*, \ldots, t_B^*$ from B bootstrap samples

Example: comparison of population means $H_0 : \mu_x = \mu_y$ vs. $H_A : \mu_x \neq \mu_y$

- if the shapes of the underlying distributions are identical, then the two distributions are the same under H_0
- choose for $\hat{F}_0$ the pooled empirical distribution function of the two samples
- the bootstrap test will be the same as the permutation test, except that random permutations will be replaced by random samples of size $n_x + n_y$ drawn **with replacement** from the pooled data

(Monte Carlo) Permutation vs. Bootstrap Resampling

- In MC sampling one samples values of the test statistic from its underlying permutation distribution
- In Bootstrapping there are two sources of error:
  1. Error caused by resampling from an empirical cumulative distribution function formed from the initial data set.
  2. Error caused from by carrying out only a finite number of resamples.
- For messier problems when the test statistic has a complicated analytically intractible distribution the bootstrap can provide a reasonable answer while the permutation test may not work.
- Permutation methods only apply in a narrow range of problems. When they apply, as in testing $F = G$ in two-sample problems, they give "exact" answers without parametric assumptions

(Monte Carlo) Permutation vs. Bootstrap Resampling (cont)

An example comparing the location of two distributions by one-sided tests:

```
x <- scan()
16 23 38 94 99 141 197

y <- scan()
10 27 31 40 46 50 52 104 146
```

The observed test statistic $\bar{X} - \bar{Y}$ is

```
mean(x)-mean(y)
```

```
[1] 30.63492
```

(Monte Carlo) Permutation vs. Bootstrap Resampling (cont)

We want to compute $P(\bar{X} - \bar{Y} \geq 30.63 | F = G)$. The permutation test is done using $16!/(7!9!) = 11440$ partitions of the 16 cases into two groups of 9 and 7, respectively.

```
library(exactRankTests)
perm.test(x, y, alter="greater")

        2-sample Permutation Test

data:  x and y
T = 608, p-value = 0.1406
alternative hypothesis: true mu is greater than 0
```

A bootstrap test was done with 1000 bootstrap samples.
In 122 of these the bootstrap estimate of $\bar{X} - \bar{Y}$ equalled or exceeded the original mean difference of 30.63.
Thus the bootstrap estimate of the p-value is $122/10000 = 0.122$

When does the permutation test fail?

The permutation test is exact, if:

- in the one-sample problem, the variables have a symmetric distribution
- in the two- and k-sample problem, the variables are exchangeable among the samples

A permutation test for comparing the means of two populations does not fail, if either the variances are the same, or the sample sizes are the same (cp. Romano, JASA 1990, p.686-692).

A permutation test for the difference of the medians of two distributions will not be exact, even asymptotically, unless the underlying distributions are the same. This is independent of the sample sizes (cp. Romano, JASA 1990, p.686-692).

A permutation test fails, if one tests for interaction in an unbalanced design! (cp. Good, Permutation Tests, 1993).

When does the permutation test fail?

An example comparing the location of two distributions by two-sided tests, where the true means and variances as well as the two group sizes are different:

```
x <- rnorm(25,0,1)
y <- rnorm(75,1,4)

perm.test(x,y,exact=T)

boot.test(x,y,B=5000)

t.test(x,y)
```

Two-sample permutation test: $p = 0.174$
Two-sample bootstrap test using the t-test statistic: $p = 0.0302$
Welch two-sample t-test: $p = 0.0243$

Bootstrapping in more complicated situations

- Two different ways to bootstrap a regression model
  1. Bootstrap data pairs $x_i = (c_i, y_i)$
  2. Bootstrap the residuals

$$\Rightarrow \quad x_i = (c_i, c_i\hat{\beta} + \hat{\varepsilon}_{i_1})$$

Example

Regression model $y_i = c_i\beta + \varepsilon_i$

- To generate x^* we first select a random sample of bootstrap error terms

$$\hat{F} \rightarrow \varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)$$

- Bootstrap responses y_i^* are generated by

$$y_i^* = c_i\hat{\beta} + \varepsilon_i^* \qquad \hat{\beta} \text{ fixed}$$

- The parametric bootstrap data set is then $x^* = (x_1^*, ..., x_n^*)$ with $x_i^* = (c_i, y_i^*)$
- The bootstrap least squares estimator $\hat{\beta}^*$ is

$$\hat{\beta}^* = (C^T C)^{-1} C^T y^*$$

with

$$var(\hat{\beta}^*) = \hat{\sigma}_F^2 (C^T C)^{-1}$$

Notes:

- The parametric bootstrap test samples new data under the Null hypothesis by assuming a distribution of the test statistic which depend upon nuisance parameters.
- Bootstrapping pairs is less sensitive to assumptions than bootstrapping residuals
- Bootstrap confidence intervals can be calculated using the quantiles of the calculated bootstrap sample.

When might the bootstrap fail?

- Incomplete data
- Dependent data
- Dirty data ("outliers")

Outlook: Bootstrap aggregating

Bagging (acronym for Bootstrap aggregating)
fits many (large) trees to bootstrap-resampled versions of the training data, and builds classifiers/predictors by majority vote/average.

Growing the tree on a learning sample $\mathcal{L} = \{(y_i, x_i), i = 1, ..., n\}$ provides a predictor $\phi(x, \mathcal{L})$ for the response y.

Using a sequence of learning samples $\{\mathcal{L}_k\}$, drawn from the same underlying distribution as $\mathcal{L}$, allows to replace $\phi(x, \mathcal{L})$ by the average of $\phi(x, \mathcal{L}_k)$,

$$\phi_A(x) \;=\; E\phi(x, \mathcal{L})$$

Bootstrapping lets us imitate the process leading to $\phi_A(x)$ by using B bootstrap samples $\mathcal{L}_1, ..., \mathcal{L}_B$ from $\mathcal{L}$ and computing

$$\phi_B(x) \;=\; ave_{b=1}^{B}\phi(x, \mathcal{L}_b)$$

Bagging Survival Estimates

Possibilities:

1. Pointwise median of Kaplan-Meier curves
2. Do not aggregate point predictions but predict the conditional survival probability function by bagged survival trees:
   - Draw B bootstrap samples
   - Construct survival tree on each bootstrap sample
   - For a new observation
     – determine the B terminal leaves corresponding to this observation
     – aggregate the B bootstrap subgroups of observations belonging to these B terminal leaves, i.e. patients are aggregated with repetition.
     – compute the Kaplan-Meier estimate using the aggregated set of observations