Bradley Efron

Statistical Data Analysis

In the Computer Age

B Efron & R. Tibshirani

"Science" July 26 1991
p390-395

"Today's Data analyst can afford to expend more computation on a single problem than the world's yearly total of statistical computation in the 1920's."

n = 15 Observed Lifetimes:

.143	.182	.256	.260	.270	.437
X_1	X_2	X_3	X_4	X_5	X_6

.509	.611	.712	1.04	1.09	1.15
X_7	X_8	X_9	X_{10}	X_{11}	X_{12}

1.46	1.88	2.08
X_{13}	X_{14}	X_{15}

Sample Mean

$\bar{x} = .804$

Sample Median

$\hat{\theta} = .611$

Estimated Standard Error

$\hat{\sigma} = .155$

$\hat{\sigma} = [\sum (x_i - \bar{x})^2 / n^2]^{\frac{1}{2}}$

$\hat{\sigma} = ??$

We have

DATA $\underset{\sim}{y}$

e.g. $\underset{\sim}{y} = (x_1, x_2, \cdots, x_{15})$ random sample of 15 lifetimes

PARAMETER Of Interest θ

e.g. the true expected lifetime
or true median lifetime.

We Want to estimate

θ from $\underset{\sim}{y}$

Two Basic Questions

Question 1: What statistic $\hat{\theta}(\underset{\sim}{y})$ should we use to estimate θ?

Question 2: How accurate is $\hat{\theta}$ as an estimate of θ?

Maximum Likelihood Theory

Answer 1: Use $\hat{\theta}(\underset{\sim}{y})$ the MLE

Answer 2: Standard Error of $\hat{\theta}$ is approximately

$$\hat{\sigma} = \frac{1}{\sqrt{\text{Fisher Info}}}$$

BOOTSTRAP is a more general way to answer Q2.

- Less Parametric Modelling
  (even nonparametric)
- More Computation
  (x100 or 1000)
- Automatic
  (Algorithm)

The Simplest Situation

$F \xrightarrow{\text{Random Sample}} (x_1, x_2, \cdots x_n) = \underset{\sim}{y}$

(F: true distribution)

Statistic of Interest: $\bar{X} = \sum_1^n x_i / n$

$\hat{\theta}(\underset{\sim}{y})$

Simple formula for standard error:

$$\sigma(F) = [\mu_2(F)/n]^{1/2} \quad \bigstar$$

Where

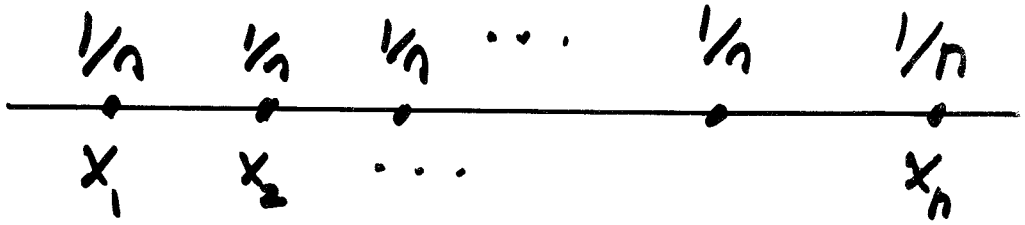
$\mu_2(F) = 2^{nd}$ Central Moment of F

$= E_F[X - E_F[X]]^2$

= Variance of one observation "X"

Estimating $\sigma(F)$

Empirical Distribution: $\hat{F}$:



$$\mu_2(\hat{F}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n$$

Gives <u>estimated</u> standard error

$$\hat{\sigma}(\hat{F}) = [\mu_2(\hat{F})/n]^{1/2} \quad \bigstar$$

$\hat{F}$ for F

$$= \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 / n^2\right]^{1/2}$$

a nice simple formula.

For More Complicated Statistics

e.g. $\hat{\theta}(\underset{\sim}{y})$ = Sample Median

- No Simple Formula for
  $\sigma(F)$ = Standard Error of $\hat{\theta}(\underset{\sim}{y})$

- Can't "Substitute $\hat{F}$ for F"

- **Bootstrap**: computer algorithm for finding numerical value of

$$\sigma(\hat{F}) \equiv \hat{\sigma}$$

Empirical ↗ ↖ Bootstrap Estimate of Standard error

Bootstrap Sampling

"$\hat{F} \rightarrow \underset{\sim}{y}^* = (x_1^*, x_2^*, \cdots x_n^*)$"

Means that you

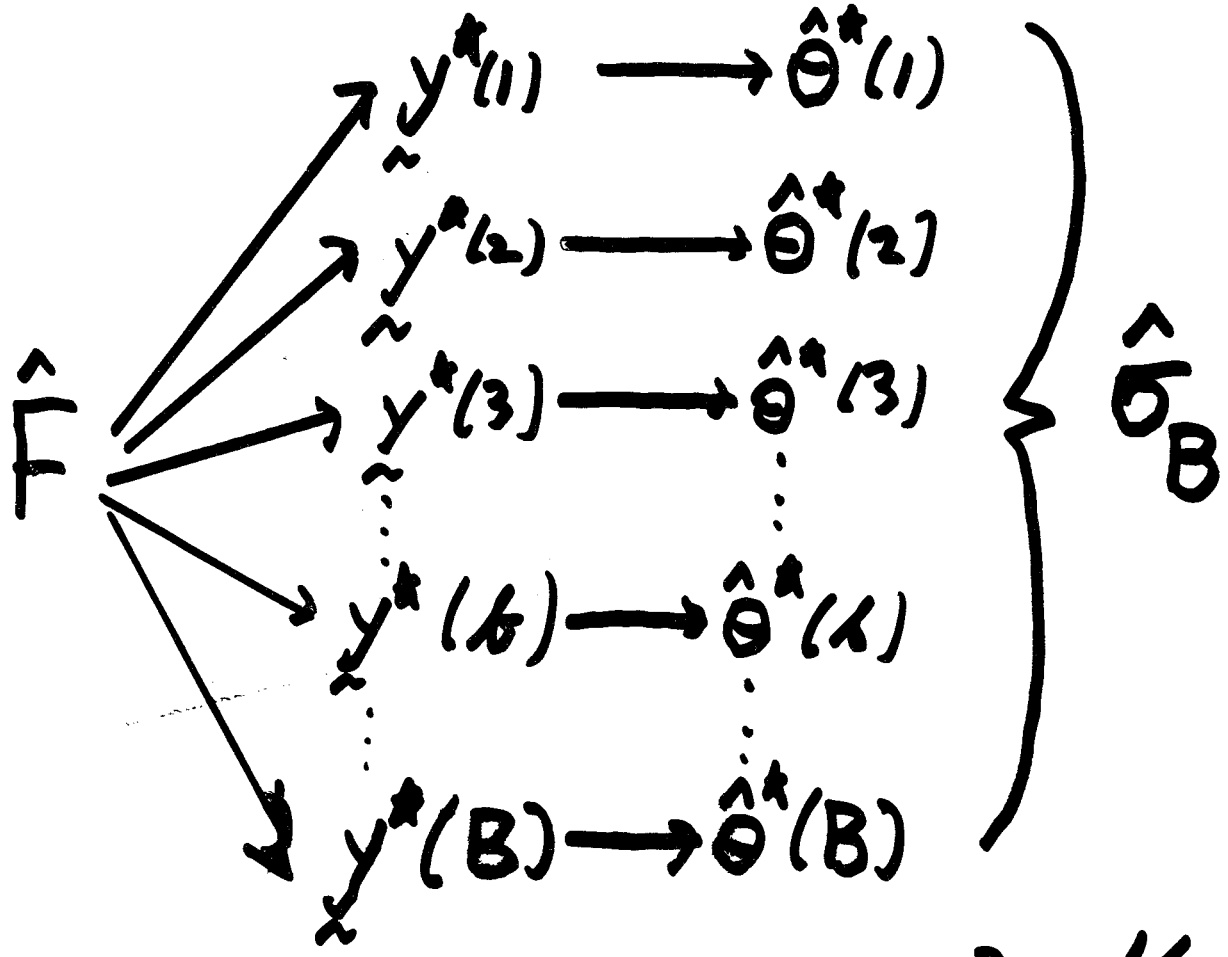
"Draw a random sample of size n from $\hat{F}$"

Equivalent: $\underset{\sim}{y}^* = (x_1^*, x_2^*, \cdots x_n^*)$ is a random sample of size n drawn with replacement from

$$\{x_1, x_2, \cdots, x_n\}$$

Definition: $\underset{\sim}{y}^*$ is a Bootstrap Sample

BOOTSTRAP ALGORITHM



$$\hat{\sigma}_B = \left\{ \frac{\sum [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1} \right\}^{1/2}$$

As $B \to \infty$, $\hat{\sigma}_B \to \hat{\sigma}_\infty = \sigma(\hat{F})$,

the bootstrap estimate of standard error.

Results for the 15 lifetimes

B = 100 Bootstraps

Sample Mean $\hat{\sigma}_{100} = .156$ → $\hat{\sigma}_{MLE} = .155$

$\left\{ \Sigma (x_i - \bar{x})^2 / n \right\}^{1/2}$

Sample Median $\hat{\sigma}_{100} = .229$

"boot (y, 100, "median")"

pk

Usually $B = 100$ is plenty*:

Lifetime Data, $\hat{\theta}$ = Sample Median

B:	25	50	100	200	1000
$\hat{\sigma}_B$:	.23	.22	.23	.25	.25

* For Standard Errors

MORAL

- Most of the familiar methods of getting standard errors are approx. versions of $\hat{\sigma} = \sigma(\hat{F})$.

- Bootstrap evaluates $\sigma(\hat{F})$ directly by brute force Monte Carlo

- Mathematical complications of step $\underset{\sim}{y} \to \hat{\theta}$ don't bother the bootstrap.

Ellipticity Data

- 51 Pairs $x_i = (y_i, z_i)$ $i = 1, 2 \cdots 51$
- y_i = temperature z_i = ellipticity
- Model:

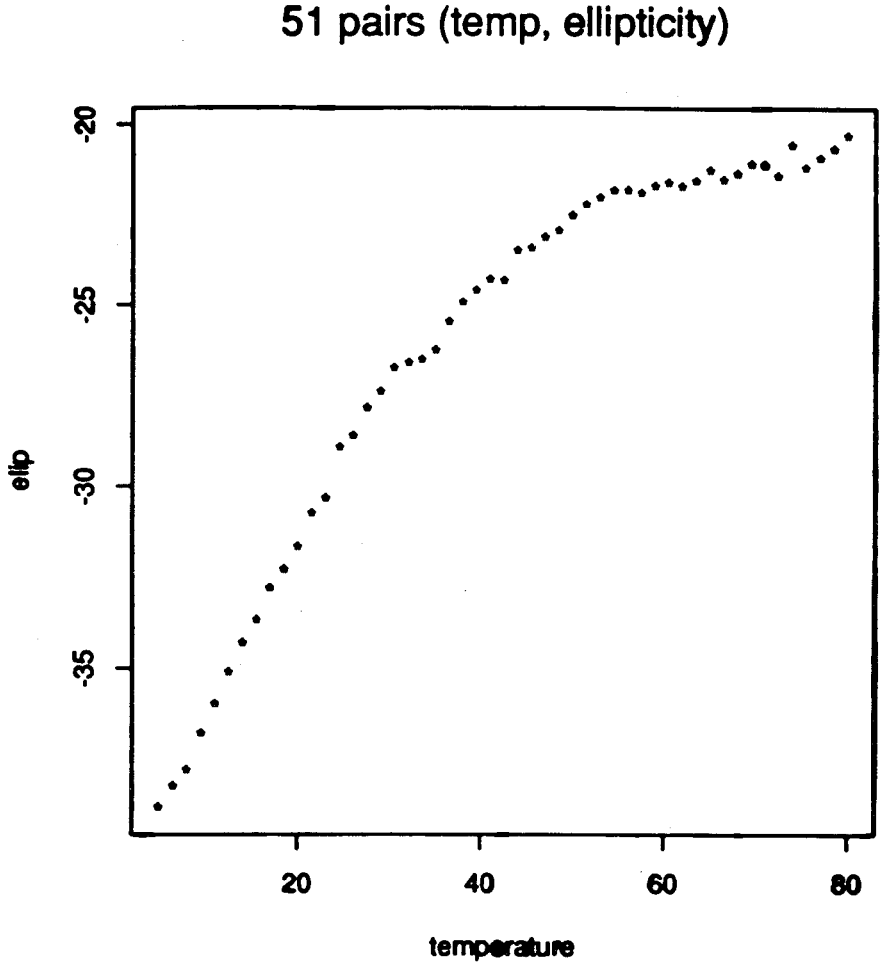
$$z_0 = \frac{a + b \cdot e^u}{1 + e^u}$$

where

$$u = d \cdot \left[\frac{1}{c} - \frac{1}{x + 273.15} \right]$$

- Nonlinear model with four parameters (a, b, c, d)
- Fit by least squares:

$$\min_{(a,b,c,d)} \sum_{i=1}^{51} [z_i - z_{0i}]^2$$

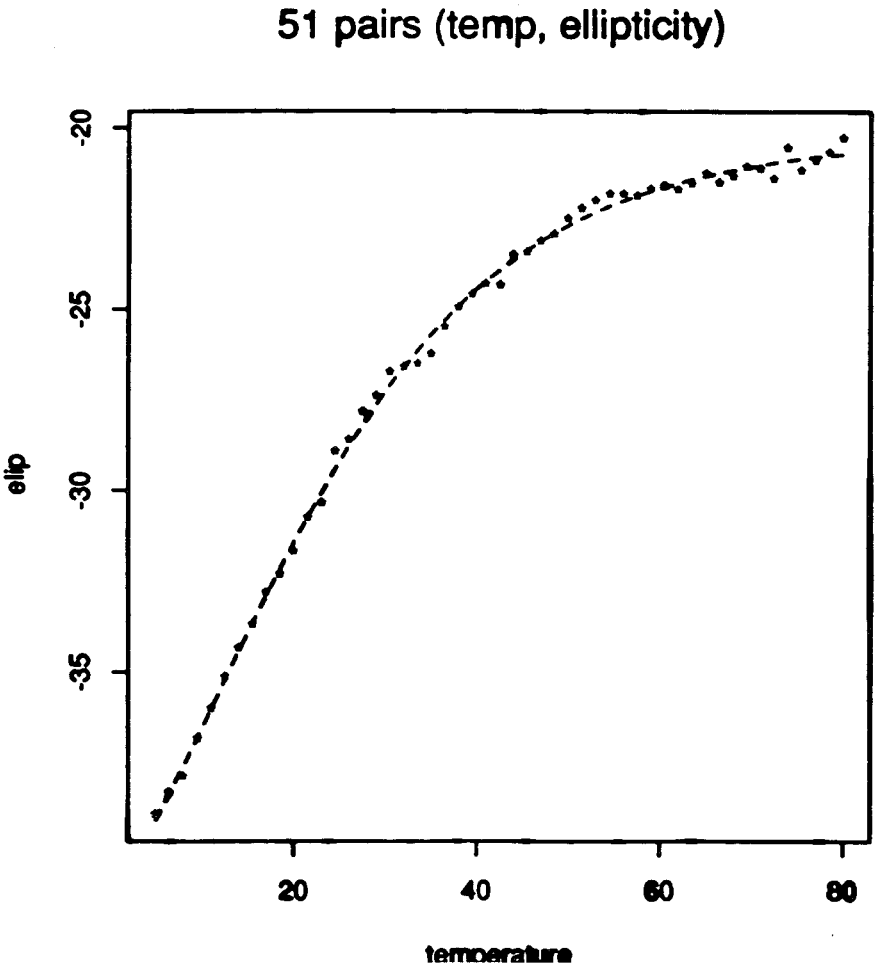


$\hat{a} = -48.3$

$\hat{b} = -20.1$

$\hat{c} = 287.5$

$\hat{d} = 6014$

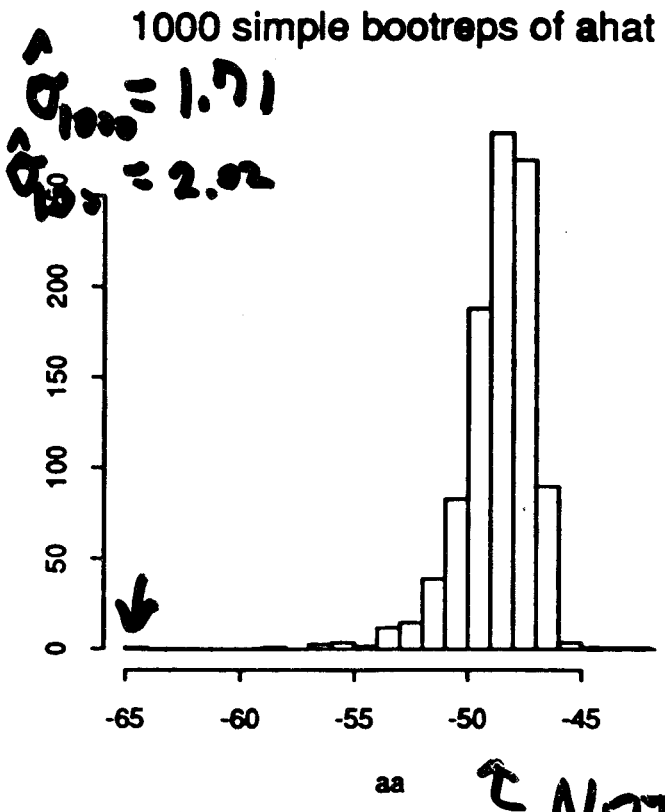


Bootstrap Analysis of $(\hat{a}, \hat{b}, \hat{c}, \hat{d})$

- Bootstrap sample $\underset{\sim}{y}^* = (x_1^*, x_2^* \ldots x_{51}^*)$ a random sample, with replacement, of 51 pairs from $\underset{\sim}{y} = (x_1, x_2, \ldots x_{51})$.

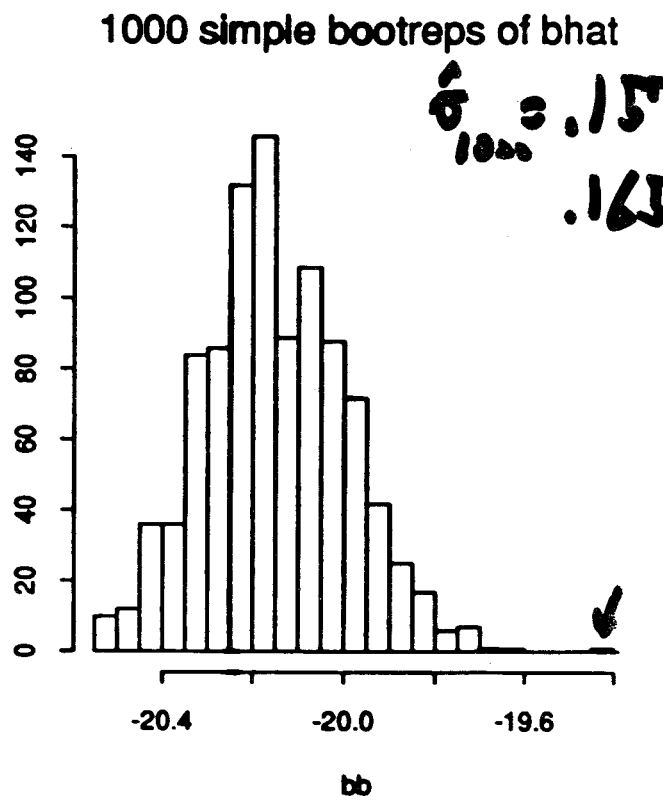
- Boot replication of $(\hat{a}, \hat{b}, \hat{c}, \hat{d})$ is $(\hat{a}^*, \hat{b}^*, \hat{c}^*, \hat{d}^*)$ the non-linear least-squares fit to the bootstrap sample $\min \sum_{i=1}^{51} [z_i^* - \hat{z}_{0i}^*]^2$

- Did $B = 100$ and $B = 1000$ Bootstraps



$\hat{\sigma}_{1000} = 1.71$

$\hat{\sigma}_{1000} = 2.02$



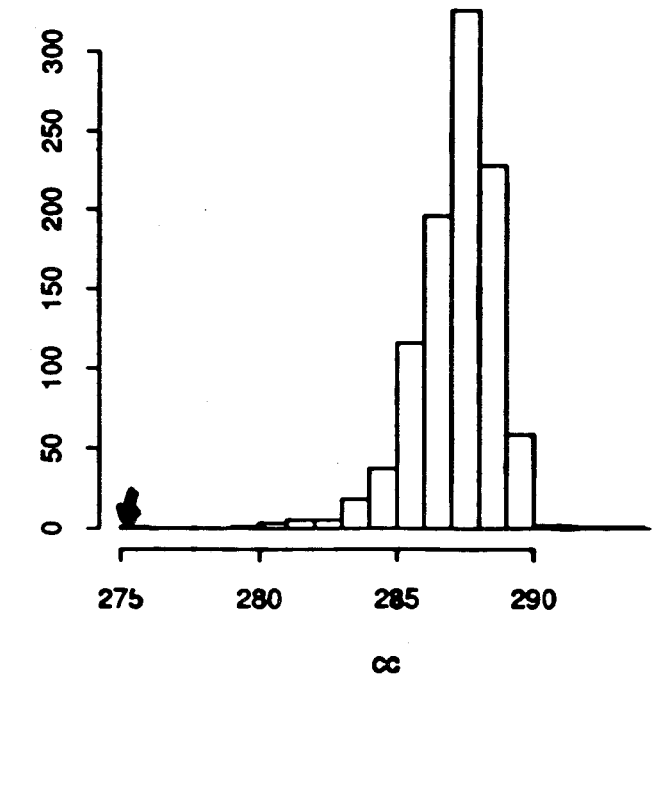
$\hat{\sigma}_{1000} = .157$

.163

NOT BELL-SHAPED!

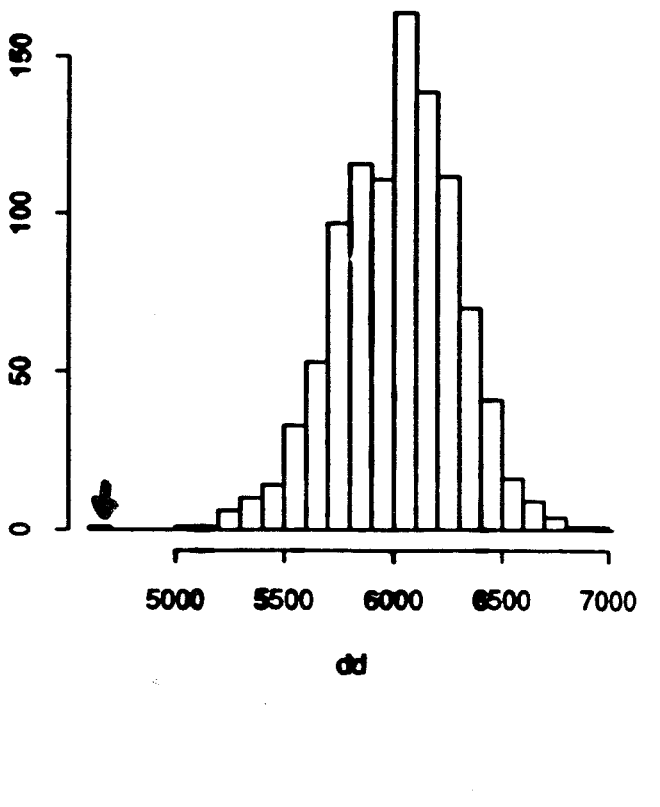
1.77

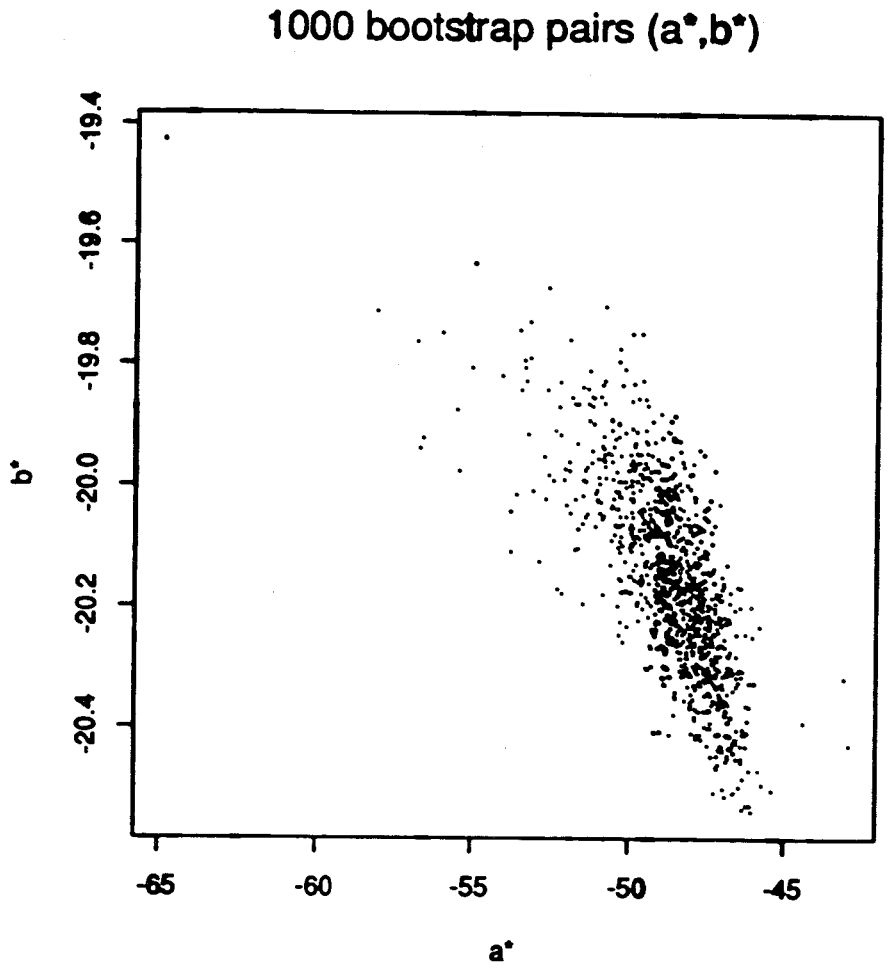
$\hat{\sigma}_{1000} = 1.52$

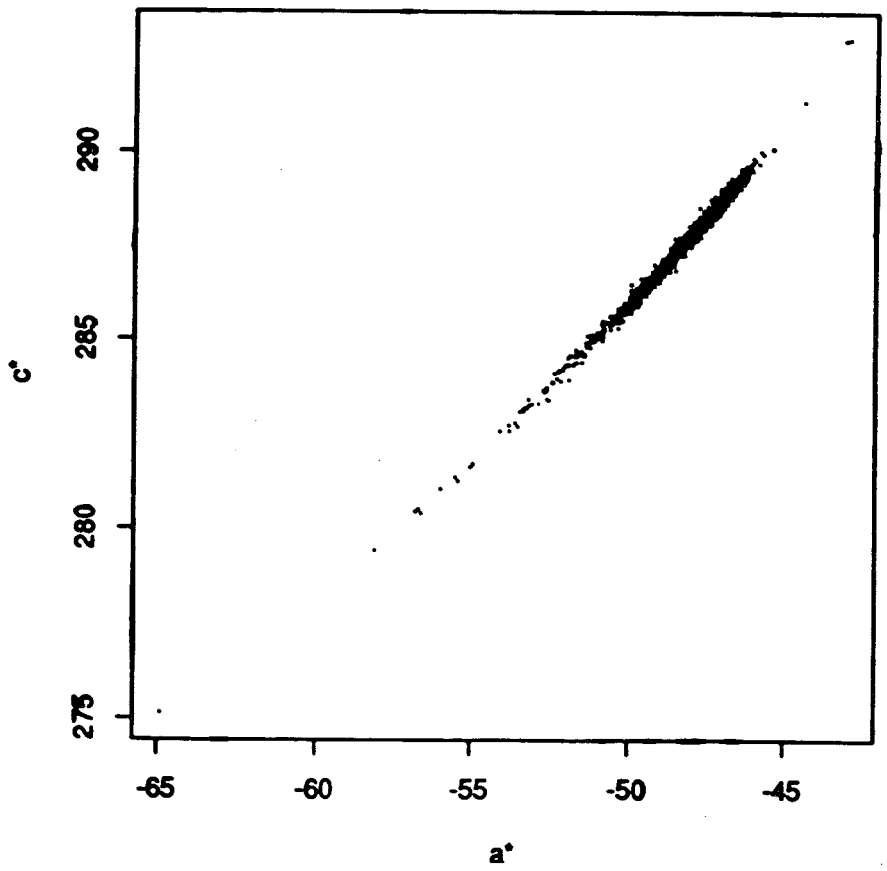


308

$\hat{\sigma}_{1000} = 277$




1000 bootstrap pairs (a*,b*)
b*
-19.4
-19.6
-19.8
-20.0
-20.2
-20.4
-65
-60
-55
-50
-45
a*


1000 bootstrap pairs (a*,c*)
c*
290
285
280
275
-65
-60
-55
-50
-45
a*

Nonparametric Regression: "Lowess"

- $n = 164$ men, $x_i = (\text{compliance}_i, \text{Chol. Decrease}_i)$

- Bootstrap:

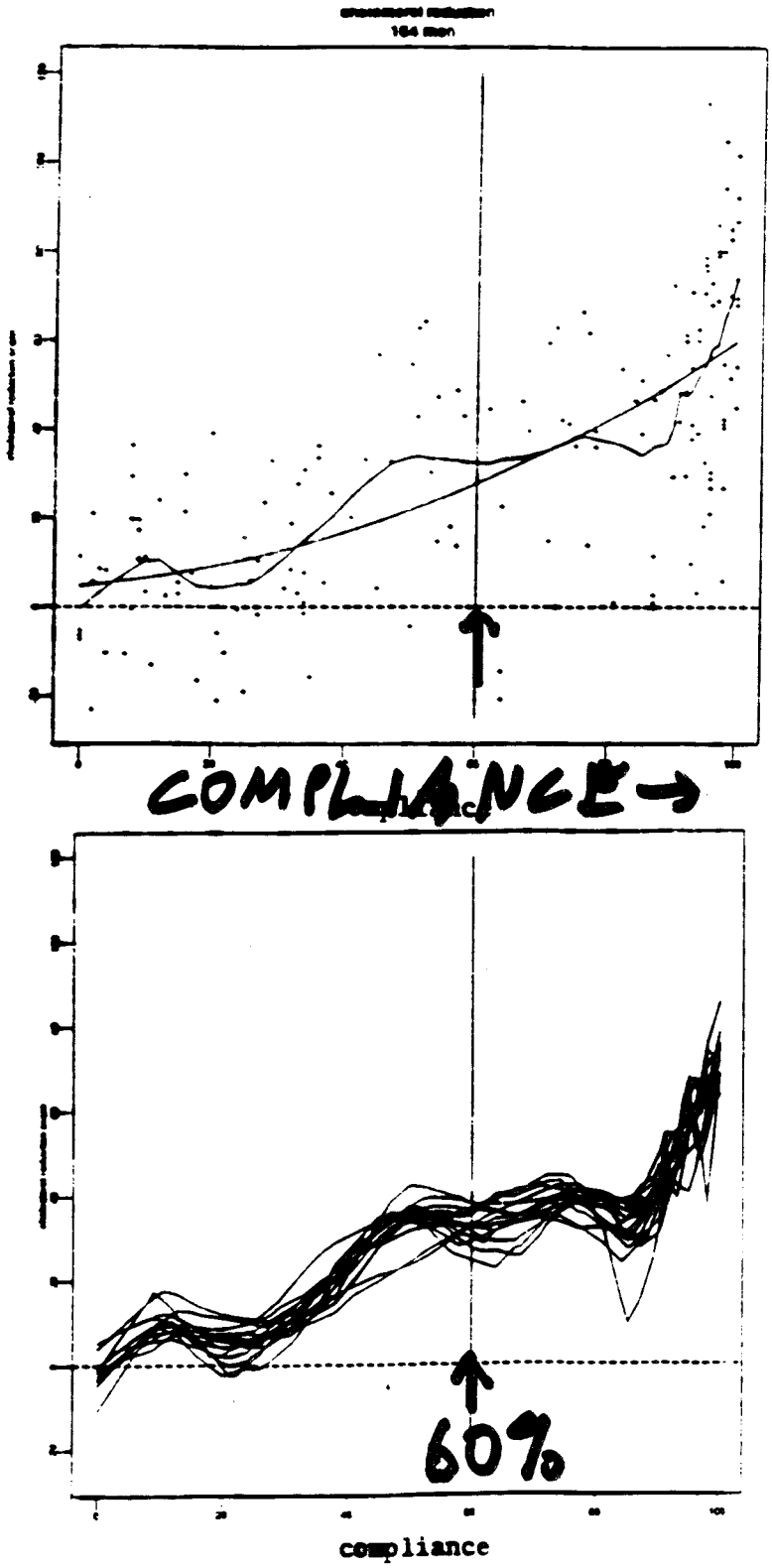
$\underset{\sim}{y}^* = (x_1^*, x_2^*, \ldots, x_{164}^*)$

- Lowess*

- $B = 50$

- $\hat{\sigma}(\text{Lowess}_{60\%}) = 5.71$

- $\hat{\sigma}(\text{Quad}) = 3.08$



"The history of science exhibits a steady tendency to eliminate intellectual effort in the solution of individual problems, by developing comprehensive formulas which can resolve by rote a whole class of them."

... Ernest Nagel, 1955