

13-6-2025

Bootstrap Example (Διόρθωση μεροληψίας)

Έστω δείγμα $x_1, \dots, x_n$ (πραγματικό)
(από κατανομή F)

Θέλουμε να εκτιμήσουμε παράμετρο θ

Έχουμε μια στατιστική συνάρτηση - εκτιμήτρια του θ

$$\hat{\theta} = \underbrace{T(x)}$$

$\hat{\theta}$ εκτίμηση του θ από το δείγμα (σημιακή)

Θέλουμε να εκτιμήσουμε τη μεροληψία
αυτής της εκτιμήτριας (T)

$$\text{Bias} = b = \underbrace{E(\hat{T})}_{\text{απόδοση}} - \underbrace{\theta}_{\text{απόδοση}}$$

$$\hat{b} = \underbrace{\text{Εκτίμηση του } E(T) \text{ μέσω bootstrap}}_{\text{bootstrapped}} - \hat{\theta}$$

Εκτίμηση μέσω bootstrap.

N Δείγματα bootstrap (sample με επανάληψη
από το αρχικό δείγμα
 $x_1, \dots, x_n$)

Γε κάθε δείγμα υπολογίζω το $T(x) \rightarrow T_1, T_2, \dots, T_N$

Εκτίμηση του $E(T) = \overline{T_N} = \frac{T_1 + \dots + T_N}{N}$

$$\Rightarrow \hat{b} = \overline{T_N} - \hat{\theta}$$

Εστω π.χ. $\hat{\theta} = 50$ $\overline{T_N} = 40 \Rightarrow \hat{b} = -10$

Διόρθωση του $\hat{\theta}$ για να λάβουμε
υπόψη των αμεροληψιών:

$$\tilde{\theta} = \hat{\theta} - \hat{b} = \hat{\theta} - (\overline{T_N} - \hat{\theta})$$

bias corrected estimate $\left[\tilde{\theta} = 2\hat{\theta} - \overline{T_N} \right]$

Εφαρμογή στον Εκτίμηση του λ της $\text{Exp}(\lambda)$.

$$X \sim \text{Exp}(\lambda) \quad (\theta: \lambda = \frac{1}{E(X)})$$

$$T(x) = \frac{1}{\overline{X_n}} \quad \text{δείγμα}(x_1, \dots, x_n)$$

Εστω δειγμά από την $\text{Exp}(\lambda)$ με άγνωστο λ .

$$(x_1, \dots, x_n) \leftarrow \text{το πρώτο δειγμά}$$

$$\hat{\theta} = \hat{\lambda} = \frac{1}{\bar{x}_n} = \frac{1}{\frac{x_1 + \dots + x_n}{n}}$$

Παίρνουμε N bootstrap samples από το $(x_1, \dots, x_n)$
Για καθένα υπολ. $T_j(x) = \frac{1}{\bar{x}_n}$, $j=1, \dots, N$

$$\Rightarrow \bar{T}_N = \frac{T_1 + \dots + T_N}{N}$$

bias correction $\tilde{\lambda} = 2\hat{\lambda} - \bar{T}_N = \frac{2}{\bar{x}_n} - \bar{T}_N$

Υποδοίον για $\lambda=2$ ($EX=1/2$), $n=10$

- 1) Δημιουργώ δείγμα μεγέθους $n=10$ από $\text{Exp}(2)$
- 2) Εφαρμόζω τα παραπάνω σε αυτό το δείγμα.

Συσχετισμένα δεδομένα

(X, Y)

οχι ανεξάρτητες

Ιδέα Έστω ότι γνωρίζουμε $F_x(x)$:
ληϊδωρία

και τη δεσμευμένη Y/X : $F_{Y/X}(y/x)$

Τότε μπορούμε να έχουμε μια γέννηση από $F_x(x)$

Διευρυνόμε μια παρατήρηση x από F_x

κ' έπειτα μια παρατήρηση y από $F(y/x)$

Παράδειγμα

Στο γραμμικό μοντέλο $Y = b_0 + b_1 x + \varepsilon$

$$\varepsilon \sim N(0, \sigma^2)$$

οι εκτιμήσεις LSE $\hat{b}_0, \hat{b}_1$: αμερόμητες

Πως μπορούμε να το επαμειώσουμε
μέσω προσηλωσης?

Πείραμα προσομοίωσης

Παίρνουμε συγκεκριμένο διάνυσμα x (π.χ. age)
 $x = (10, 15, 20, \dots, 100)$

Υποθέτω ένα γραμμικό μοντέλο

$$Y = b_0 + b_1 x + \varepsilon$$

όπου $b_0 = 30, b_1 = 1, \sigma = 30$

Δημιουργούμε N προσομ. δείγματα Y μεγέθους (x)

σε καθένα $Y_j = b_0 + b_1 x_j + \varepsilon$, $j=1, \dots, n$
 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

από το $(Y, x) \Rightarrow \boxed{\hat{b}_0, \hat{b}_1, \hat{\sigma}^2 = \text{MSE}}$

Παίρνουμε πίσω πάλι το $\hat{b}_0, \hat{b}_1, \hat{\sigma}^2$ ($? \approx b_0, b_1$)

Προσομ.
επανάληψη

1

2

⋮

⋮

N

$\hat{b}_0$

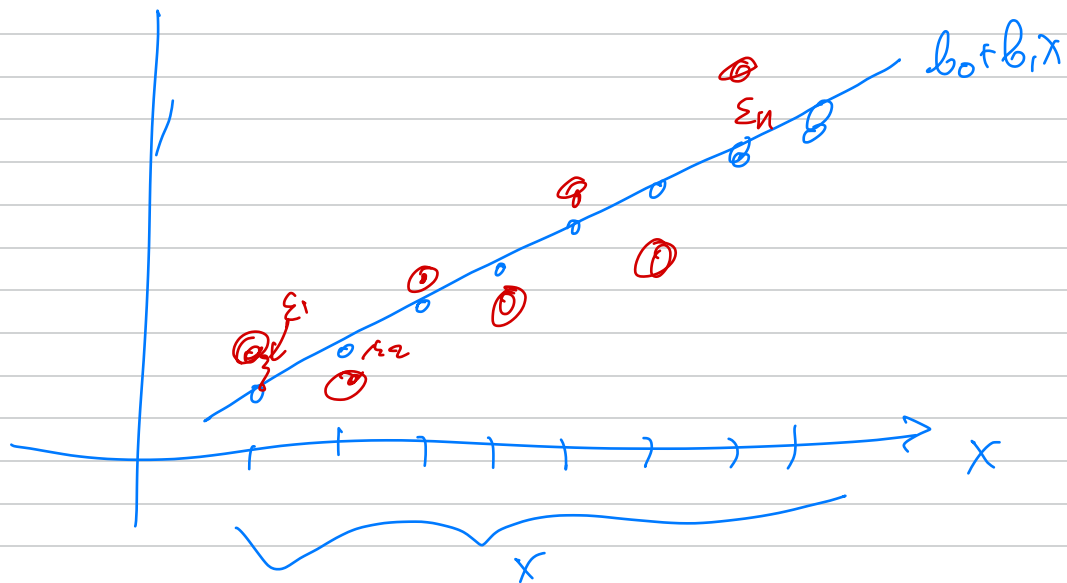
$\hat{b}_1$

MSE

$N \times 3$

results

results = matrix(rep(0, 3xN), N, 3)



Παράδειγμα 2

$$Y \in \{0, 1\}$$

$$P(Y=1|x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} \quad \Leftrightarrow$$

$$\Rightarrow \log \frac{P(Y=1|x)}{P(Y=0|x)} = b_0 + b_1 x$$

$$\text{Qu } p(x) = P(Y=1|x) \Rightarrow \log \frac{p(x)}{1-p(x)} = b_0 + b_1 x$$

$\hat{b}_0, \hat{b}_1$ MLE (max-likelihood estimates)

1) $\hat{b}_0, \hat{b}_1$ είναι ανεξαρτητές;

2) $\hat{p}(x)$ είναι ανεξαρτητική?

Επίσημο Για να δημιουργήσουμε
simulated dataset (x, y)

με $x = \text{fixed}$.

$y|x$: $P(Y=1|x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} = p(x)$

Δεδο. x : $Y \sim \text{Bernoulli}(p(x)) \sim \text{Binomial}(1, p(x))$

R:
$$(a) \begin{cases} u = \text{runif}(1, 0, 1) \\ \text{if } u < p(x) \Rightarrow Y=1 \text{ else } Y=0. \end{cases}$$

$(b) \quad Y = \text{rbinom}(1, 1, p(x))$

Υποδοίωση

Εστω $x = 1, 2, \dots, 10$

$$\log \frac{p(x)}{1-p(x)} = b_0 + b_1 x,$$

$$\begin{aligned} b_0 &= 0.1 \\ b_1 &= 0.05 \end{aligned}$$

Αντιμετώπιση Ελλειπόντων Τιμών

x=age	y=
15	32
20	37
18	40
-	35
-	17
20	36

← missing values
(R/Stata μη αγνοούνται)

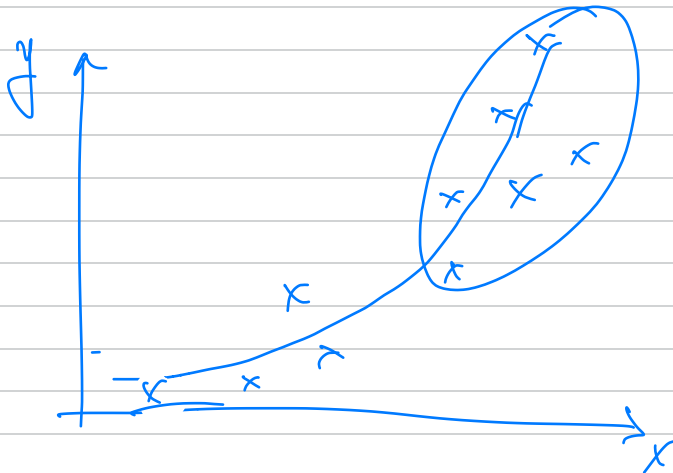
$N=6$

Ερωτήματα Γιατί λείπουν;

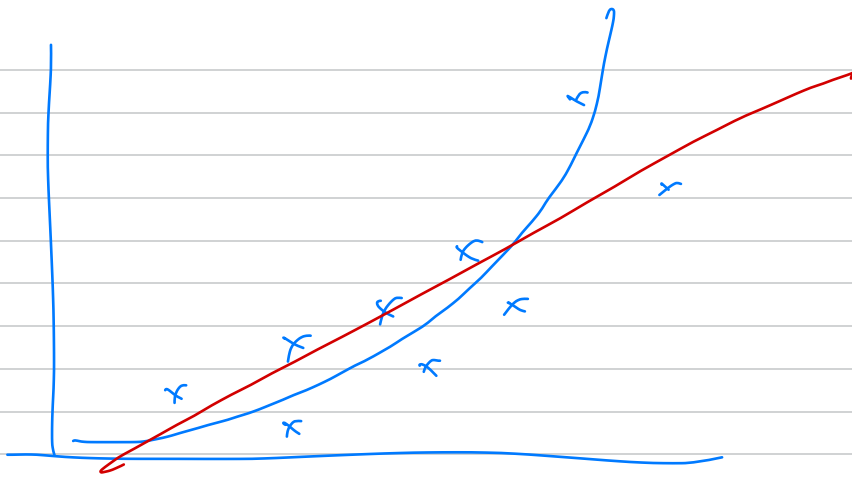
Ⓐ Missing Completely at random

Κάθε περίπτωση έχει μια πιθανότητα να γίνει ή όχι ανεξάρτητα από τις υπόλοιπες

Ⓑ Missing not at random / informative



Ερω ου για μεγάλα x η $P(\text{missing}) \approx 0.9$



① missing Completely at random Simulated dataset

On 20 aprils Scitex éter feged n
n' Defoupe 20 10% 200 nqar. 10 éter
missing

missing_index = sample(1:n, 0.1 * n, rep = F)

1, 2, 3, ..., n

n x. n = 30 0.1 n = 3

missing_index = (7, 10, 29)

x[missing_index] = na

Missing at random informative
Naodétype $q(x) = P(\text{missing} | x)$
for (j in 1:n)
pmiss = q(x[j])

$$\left(\begin{array}{l} u = \text{runif}(\phi, 0, 1) \\ \text{if } u < p_{\text{miss}} \Rightarrow x[j] = \text{na} \end{array} \right)$$

εναλλακτικά

$$\left(\begin{array}{l} y = \text{rbinom}(1, 1, p_{\text{miss}}) \\ \text{if } y == 1 \quad x[j] = \text{na} \end{array} \right)$$