Regression

17th oct 2019

Bootstrap simulation.

$X_1, \ldots, X_n \overset{iid}{\sim} f(\,)$ with $E(X) = \mu$, $V(x) = \sigma^2 < \infty$

n is small.

Interested in inference about μ.

$\bar{X} \rightarrow$ Estimation

$\rightarrow$ For n ... we need sampling distribution.

If n is large we know

$$\bar{x} \overset{approx}{\sim} N\left(\mu, \sigma^2/n\right) \quad [CLT]$$

but if n is small then $\bar{x} \sim ?$

Suppose $\bar{X}_n \sim G_n$. [It has its own dist. denoted by G_n].

$$G_n(x) = P_G(\bar{X}_n \le x).$$

equivalently, $x_1, \ldots, x_n \sim f(x)$. [cdf $F(x)$]

Claim: If we can make an educated guess of $F(x)$, the "we are done".

Detour:

Suppose, $F(x) = $ Gammer.

Don't know the sampling dist. of $\bar{X}$

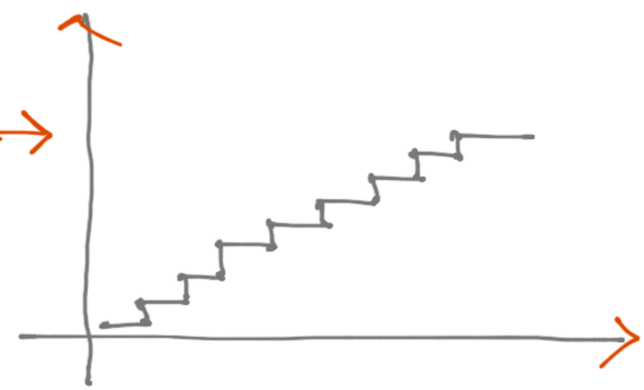
$$F(x) \longrightarrow \begin{array}{c} (x_1, \ldots x_n) \\ (x_1, \ldots x_n) \\ \vdots \end{array} \quad \boxed{\begin{array}{c} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \end{array}}$$

All we are needed to do is to make a guess of $F(x)$.

Empirical CDF.

$$F_n(x) = P_n(X \le x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \le x)$$

$$[\text{ie } (\# X_i \le x / n)]$$

[It will give a step function] 

Given $X_1 \dots X_n$ define $Y_1 \dots Y_n$ where

$$Y_i = \mathbb{I}(X_i \le x) = \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{o.w.} \end{cases}$$

if $X_1 \ldots X_n \sim F(x)$

then $Y_1 \ldots Y_n \sim \text{Bernoulli}(F(x))$

So $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq x)$ is MLE for $F(x)$

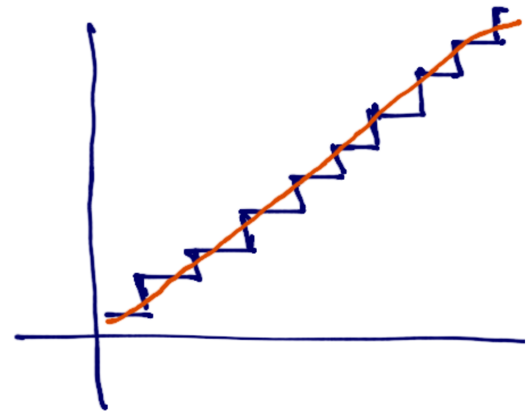
Draw SRSWR of size n from $F_n(x)$, M times.

$$\left.\begin{array}{lll} (x_{11}^*, x_{12}^*, \ldots, x_{1n}^*) \text{ from } F_n(x) & \rightarrow & \bar{X}_1 \\ (x_{21}^*, x_{22}^*, \ldots, x_{2n}^*) \text{ from } F_n(x) & \rightarrow & \bar{X}_2 \\ \vdots & & \vdots \\ (x_{M1}^*, x_{M2}^*, \ldots, x_{Mn}^*) \text{ from } F_n(x) & \rightarrow & \bar{X}_M \end{array}\right\} \text{BOOTSTRAP sample}$$

$G_n^*(x) \equiv$ Bootstrap emperical CDF of Sampling distribution of $\bar{X}$.

* This is called non-parametric Boot-strap method.

Becouse, we are not smooting



If one wonts to smooth it, one has to assume a distribution.

which makes it parametric.

Let's focus on non-parametric bootstrap.

So basically.

$$(x_1 \ldots x_n) \xrightarrow{SRSWR} \underbrace{(x^*_{i_1}, x^*_{i_2} \ldots x^*_{i_n}) \to \bar{x}^*_i}_{M \text{ times}, \quad i = 1(1)M.}$$

Gives $(\bar{x}^*_1, \ldots \bar{x}^*_M) \to$ (s.d of $\bar{x}^*_i$)

$$\text{(s.d of } \bar{x}^*_i) \approx \text{(s.e. of } \bar{x}).$$

Exercise.

$$X_1 \ldots\ldots X_{10} \overset{iid}{\sim} N(5, 2)$$

$$\bar{X}_{10} \sim N(5, 2/10)$$

$$\text{s.e of } \bar{x}_{10} = \sqrt{2/10}$$

take $M = 1000$.

From $\underset{\sim}{x}$ draw $\underset{\sim}{x}_i^*$ by SRSWR.

$$\underset{\sim}{x}_i = (x_{i1}^* \ldots x_{i10}^*) \text{ — } \bar{x}_i = \frac{1}{10}\sum_{j=1}^{10} x_{ij}^* \rightarrow \text{store}$$

Stored list gives 1000 $\bar{x}_i$.

find the s.d of the lis.

it should be approximately equal to $\sqrt{2/10}$.

It originally developed for small sample. but it works perfectly with any sample size

Briemen (2003) $\rightarrow$ Random Forest.

let dataset $D = \{(\underset{\sim}{x_i}, y_i) \mid i = 1(1)n\}$.

Draw samples $D_1^*, D_2^*, \dots D_M^*$ using some sampling strategy from D.

Build Decision Trees on $D_1^*, D_2^*, \dots D_M^*$.

let x_0 is a test point.

- Plug x_0 into each of $D_1^*, D_2^* \dots D_M^*$
- We get a prediction from each of them,

  let the predictions be $y_{01}^*, y_{02}^*, \dots y_{0M}^*$.

For classification problem:

final predicted y is $\text{mode}\{y_{01}^*, y_{02}^*, \dots y_{0M}^*\}$

i.e the label predicted most # times.

Scaling (Error margin) is not an issue here.
Unlike in case of Bootstrap sampling we are required to draw samples of same size as that of observed.

In case of Random Forest:

let the data Matrix be, $D = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ \vdots & & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{bmatrix}$

Now, → Draw M random instances (or rows)

→ Dont take all p features, draw q many features randomly. (obviously $q < p$).