Predictive Analytics
Regression and Classification

Lecture 4 : Part 2

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2020

Bootstrap Statistics

- Bootstap statistics is an algorithmic strategy, which typically resort to SRSWR scheme
- It falls under the braoder class of resampling strategy.
- Bootstrap was introduced by Brad Efron (1979). The idea though apparently simple revolutionized statistics by its ability to replace analytical derivation by brute computing force.

Bootstrap Statistics

- Suppose $\{Y_1, Y_2, \cdots, Y_n\}$ are iid observations with cdf $F()$ and $T_n = T_n(Y_1, Y_2, \cdots, Y_n)$ is a statistic which estimates a parameter θ.
- The sampling distribution of T_n would depend on $F(\cdot)$
- The bootstrap idea in its simplest form is to estimate the cdf $F(\cdot)$ by empirical cdf $F_n(\cdot)$.

Result The empirical cdf $F_n(\cdot)$ is the non-parametric MLE of cdf $F(\cdot)$.

- Bootstraping based on $F_n(\cdot)$ is called nonparametric bootstrap.

Bootstrap Statistics

- Suppose $\{Y_1, Y_2, \cdots, Y_n\}$ are iid observations with cdf $F()$ and $T_n = T_n(Y_1, Y_2, \cdots, Y_n)$ is a statistic which estimates a parameter θ.

Result The empirical cdf $F_n(\cdot)$ is the non-parametric MLE of cdf $F(\cdot)$.

- We can draw sample from $F_n(\cdot)$.
- Drawing sample from $F_n(\cdot)$ is same as draw iid samples from $\{Y_1, Y_2, \cdots, Y_n\}$
- That is draw resamples from $\{Y_1, Y_2, \cdots, Y_n\}$.
- Hence we can draw as many times as we want.

Bootstrap Framework

- $\mathbf{Y}_n = \{Y_1, Y_2, \cdots, Y_n\}$ are iid random samples from $F(\cdot)$.
- $T_n = T_n(\mathbf{Y}_n)$ is a statistic for parameter θ
- Since $F(\cdot)$ is unknown. We don't know that sampling distribution of T_n.
- Hence we don't know the variance of T_n, i.e., $Var(T_n)$ and confidence interval of T_n, i.e., $CI(T_n)$.
- Resample $\mathbf{Y}^*_{nb} = \{Y^*_1, Y^*_2, \cdots, Y^*_n\}_b$ from $\mathbf{Y}_n$ using SRSWR scheme; $b = 1, 2, \cdots, B$
- For each resample b, we can compute T^*_{nb}; $b = 1, 2, \cdots, B$

Bootstrap Framework

- $\mathbf{Y}_n = \{Y_1, Y_2, \cdots, Y_n\}$ are iid random samples from $F(\cdot)$.
- $T_n = T_n(\mathbf{Y}_n)$ is a statistic for parameter θ
- Since $F(\cdot)$ is unknown. We don't know that sampling distribution of T_n.
- Hence we don't know the variance of T_n, i.e., $Var(T_n)$ and confidence interval of T_n, i.e., $CI(T_n)$.

Bootstrap Framework

- Resample $\mathbf{Y}^*_{nb} = \{Y^*_1, Y^*_2, \cdots, Y^*_n\}_b$ from $\mathbf{Y}_n$ using SRSWR scheme; $b = 1, 2, \cdots, B$
- For each resample b, we can compute T^*_{nb}; $b = 1, 2, \cdots, B$
- We can compute:

$$\bar{T}^B_n = \frac{1}{B}\sum_{b=1}^{B} T^*_{nb}; \qquad Var(T_n)^B = \frac{1}{B}\sum_{b=1}^{B}(T^*_{nb} - \bar{T}^B_n)^2$$

$$CI(T_n)^B = \{T_n + G_B^{-1}(\alpha/2)\sqrt{Var(T_n)^B}, \quad T_n + G_B^{-1}(1-\alpha/2)\sqrt{Var(T_n)^B}\},$$

where $\frac{T^*_{nb} - T_n}{\sqrt{Var(T_n)^B}} \sim G_B$.

Bootstrap Framework

- Due to SLLN, one can show, as $B \longrightarrow \infty$

$$
\begin{aligned}
\bar{T}_n^B &\longrightarrow T_n \text{ almost surely;} \\
Var(T_n)^B &\longrightarrow Var(T_n) \text{ almost surely} \\
CI(T_n)^B &\longrightarrow CI(T_n) \text{ almost surely,}
\end{aligned}
$$

$$
G^B \longrightarrow F_{T_n}(\cdot) \quad in \quad law
$$

Bootstrap Regression

- Consider the model

$$\mathbf{y}_n = \mathbf{X}_{n\times p}\boldsymbol{\beta}_p + \boldsymbol{\epsilon}_n,$$

  where $\mathbb{E}(\boldsymbol{\epsilon}) = 0$, $\mathbb{V}ar(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$, and $\boldsymbol{\epsilon} \overset{iid}{\sim} F(\cdot)$, $F(\cdot)$ is unkniwn cdf

- OLS estimator: $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$;
  and $\mathbb{V}ar(\hat{\boldsymbol{\beta}}_n) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

- Residuals: $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n$ or $\epsilon_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_n, \quad i = 1, 2, \cdots, n.$

Residual Bootstrap Regression

- Suppose $F_n(\cdot)$ is the empirical cdf of $\boldsymbol{\epsilon}$
- $\boldsymbol{\epsilon}_b^* \overset{iid}{\sim} F_n$ (i.e., $\boldsymbol{\epsilon}_b^*$ is resampled from $\boldsymbol{\epsilon}$ using SRSWR), $b = 1, 2, \cdots, B$
- Calculate:
$$\mathbf{y}_b^* = \mathbf{X}\hat{\boldsymbol{\beta}}_n + \boldsymbol{\epsilon}_b^*$$
- Estimate resample coefficients $\hat{\boldsymbol{\beta}}_{n:b}^*$ as
$$\begin{aligned} \hat{\boldsymbol{\beta}}_{n:b}^* &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_b^* \\ &= \hat{\boldsymbol{\beta}}_n + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}_b^* \\ \mathbb{E}(\hat{\boldsymbol{\beta}}_{n:b}^*) &= \hat{\boldsymbol{\beta}}_n \end{aligned}$$
- Bootstrap Estimate: $\bar{\boldsymbol{\beta}}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\beta}}_b^*$
- Bootstrap variance: $\mathbb{V}ar(\bar{\boldsymbol{\beta}}_B) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}_b^* - \bar{\boldsymbol{\beta}}_B)^2$

Paired Bootstrap Regression

- Consider the model

$$\mathbf{y}_n = \mathbf{X}_{n\times p}\boldsymbol{\beta}_p + \boldsymbol{\epsilon}_n,$$

where $\mathbb{E}(\boldsymbol{\epsilon}) = 0$, $\mathbb{V}ar(\boldsymbol{\epsilon}) = \Sigma$, and $(y_i, \mathbf{x}_i) \overset{iid}{\sim} F(\cdot)$, $F(\cdot)$ is unkniwn cdf

- Suppose $\{(y_i^*, \mathbf{x}_i^*), i = 1, 2, ...n\}_b = \mathcal{D}_b$ are iid samples from empirical $F_n(\cdots)$, where $b = 1, 2, \cdots, B$

- The estimates of $\boldsymbol{\beta}$ from b^{th} resample:

$$\hat{\boldsymbol{\beta}}_b^* = (\mathbf{X}_b^{*T}\mathbf{X}_b^*)^{-1}\mathbf{X}_b^{*T}\mathbf{y}_b^*$$

- Bootstrap Estimate: $\bar{\boldsymbol{\beta}}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\beta}}_b^*$
- Bootstrap variance: $\mathbb{V}ar(\bar{\boldsymbol{\beta}}_B) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\beta}}_b^* - \bar{\boldsymbol{\beta}}_B)^2$

Bootstrap Regression

- If the residuals are heteroscadastic, then paired Bootstrap is still a consistent estimator.
- However in case of heteroscadastic residual; the residual Bootstrap is not consistent estimator.

Paired Bootstrap Regression

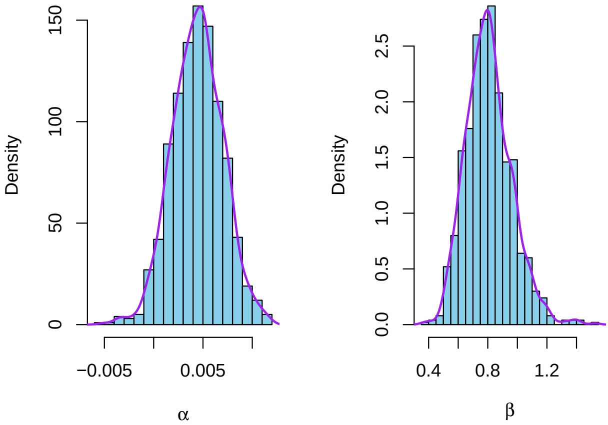
OLS Estimates of alpha and beta

	Estimate	Std. Error	t value	Pr(>|t|)
alpha	0.0046	0.0025	1.8773	0.0641
beta	0.7999	0.1556	5.1399	0.0000

Paired Bootstrap Estimates of alpha and beta

	Estimate	Std.Error	2.5%	97.5%
alpha	0.0044	0.0026	-0.0005	0.0095
beta	0.8071	0.1541	0.5379	1.1297

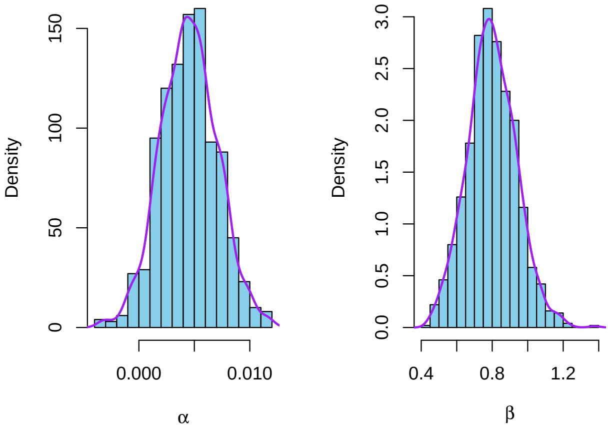
Paired Bootstrap Regression



Residual Bootstrap Regression

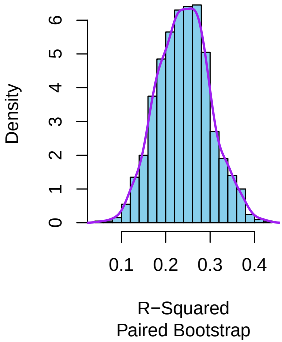
```
OLS Estimates of alpha and beta

      Estimate Std. Error t value Pr(>|t|)
alpha   0.0046     0.0025  1.8773   0.0641
beta    0.7999     0.1556  5.1399   0.0000

----------------

Residual Bootstrap Estimates of alpha and beta

      Estimate Std.Error    2.5%  97.5%
alpha   0.0045    0.0026 -0.0006 0.0097
beta    0.7982    0.1370  0.5300 1.0742

----------------

Paired Bootstrap Estimates of alpha and beta

      Estimate Std.Error    2.5%  97.5%
alpha   0.0044    0.0026 -0.0005 0.0095
beta    0.8071    0.1541  0.5379 1.1297
```

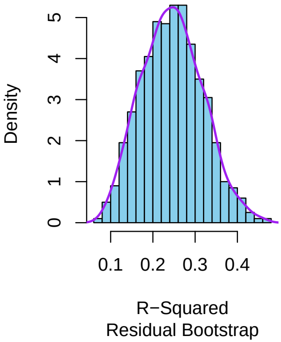
Residual Bootstrap Regression



Bootstrap Regression

```
OLS methods R-Squared =  0.239
Paired Bootstrap R-Squared CI = ( 0.128 0.366 )
Residual Bootstrap R-Squared CI =( 0.115 0.39 )
```





The idea of Bootstrap Statistics

The idea of Bootstrap Statistics or Resampling Technique can be found in

- Random Forest
- Ensamble model
- Bagging etc.

Thank you...

- Wish you a happy weekend. Stay Safe.

Thank You

sourish@cmi.ac.in