Bootstrap

Manjari Das

Learning objectives

When we are given a finite sample $\{X_1, X_2, \ldots, X_n\}$ from an unknown distribution F, Bootstrap allows us to generate multiple many independent samples from a distribution $\widehat{F}$ where $\widehat{F}$ is an approximation of F.

$$X_1, X_2, \ldots, X_n \sim F \quad \longrightarrow \quad \left.\begin{array}{c} X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)} \\ X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)} \\ \cdots \\ X_1^{(B)}, X_2^{(B)}, \ldots, X_n^{(B)} \end{array}\right\} \sim \widehat{F}.$$

- Learn the use of bootstrap method.
  - non-parametric bootstrap
  - parametric bootstrap
- Apply bootstrap to estimate a population parameter and its distribution (variance, confidence interval).
- (Optional) Apply bootstrap in linear regression.

Introductory example

Sample $X_1, X_2, \ldots, X_n \sim F$.

Suppose we want to know the population mean of $F = E(X) = \mu$.

Simple estimate is $\widehat{\mu} = \bar{X} = \frac{1}{n} \sum_i X_i$.

Question: How good an estimate is $\bar{X}$?
If we had a different sample $X_1', X_2', \ldots, X_n'$, our estimate would be $\bar{X}'$.

Question: What is the variation in the estimator?

If we had several samples from F, we could have obtained several estimates of the mean and study the variation.

This is where bootstrap lets us generate several samples when we only see a finite sample.

General problem set-up

Sample $X_1, X_2, \ldots, X_n \sim F$.

Suppose we want to know $T(F)$, some functional of F.
Examples: mean, median, variance, $\int f(x)^2 dx$, etc.

Naive approach: obtain sample estimate $\widehat{T(F)}$.
Examples:

$T(F)$	mean	median	variance	$\int f(x)^2 dx$
$\widehat{T(F)}$	$\bar{X}$	$X_{[n/2]}$	$\frac{1}{n}\sum_i (X_i - \bar{X})^2$	$\frac{1}{n}\widehat{f(X_i)}$

Obstacle: How do we get the variance and confidence interval? How good is our estimate?

Bootstrap provides a method to estimate the distribution of the functional $T(F)$.

Bootstrap method

Sample $X_1, X_2, \ldots, X_n \sim F$. $\mathcal{S} = \{X_i\}_{i=1}^{n}$

Suppose we want to know $T(F)$, some functional of F.

Bootstrap generates samples to estimate the distribution of $T(F)$.

Step 1. Draw sample $\mathcal{S}_1 \mathcal{S}_b$ from $\mathcal{S}$ with replacement.
$X_1^*, \ \ldots, \ X_n^*$.

Step 2. Evaluate $T(F)$ using $\mathcal{S}_1 \mathcal{S}_b$ to get $\widehat{T(F_b)}$.

Repeat for $b \in \{1, \ldots, B\}$

The quantities $\left\{ \widehat{T(F_1)}, \ldots, \ \widehat{T(F_B)} \right\}$ can be used to evaluate the empirical distribution of $T(F)$.

Bootstrap method

$$\begin{array}{lllll} \text{iteration 1:} & X_1^{(1)*} & X_2^{(1)*} & \dots & X_n^{(1)*} & \longrightarrow \widehat{T(F_1)} \\ \text{iteration 2:} & X_1^{(2)*} & X_2^{(2)*} & \dots & X_n^{(2)*} & \longrightarrow \widehat{T(F_2)} \\ \quad \dots & & & & & \\ \text{iteration b:} & X_1^{(b)*} & X_2^{(b)*} & \dots & X_n^{(b)*} & \longrightarrow \widehat{T(F_b)} \\ \quad \dots & & & & & \\ \text{iteration B:} & X_1^{(B)*} & X_2^{(B)*} & \dots & X_n^{(B)*} & \longrightarrow \widehat{T(F_B)} \end{array}$$

There are two ways to get the confidence interval of $T(F)$.

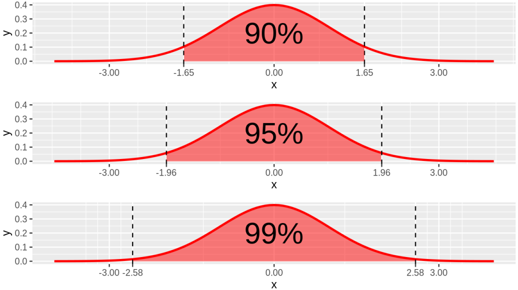
- basic $(1-\alpha)100\%$ confidence interval assuming normality $N(\textit{sample mean}, \textit{sample var}) \equiv N\left(\widehat{\mu}_{T(F)}, \widehat{\sigma}^2_{T(F)}\right)$

$$[\widehat{\mu}_{T(F)} - z_{\alpha/2}\widehat{\sigma}_{T(F)}, \ \widehat{\mu}_{T(F)} + z_{\alpha/2}\widehat{\sigma}_{T(F)}].$$

- quantile confidence interval

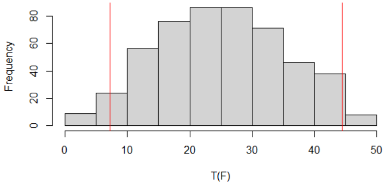
Quantile Confidence Interval

Normal(0,1) intervals, $\alpha = 0.1, 0.05, 0.01$



For a $(1-\alpha)100\%$ interval, leave out $\dfrac{100\alpha}{2}\%$ values on either tail.

Following the idea above, we can apply the "leaving out" technique on the estimates distribution of $T(F)$ as below.



$(1-\alpha)100\%$ interval is

$$\boxed{[quantile_{\alpha/2}, quantile_{1-\alpha/2}]}.$$

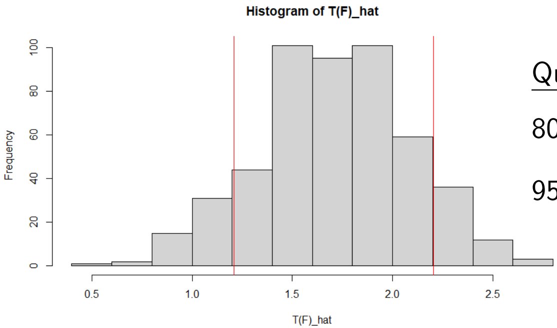
Worked out example

$\mathcal{S} = \{2.6,\ 0.89,\ 1.6,\ 2.76,\ 1.01,\ 3.06,\ 1.17\}$.
$T(F) = mean$.

b	1	2	3	4	5	6	7	$\widehat{T(F)}$
1	3.06	0.89	3.06	2.76	1.60	1.01	1.01	1.91
2	0.89	1.01	0.17	3.06	1.01	1.60	3.06	1.54
3	0.89	1.01	3.06	2.76	3.06	0.89	1.60	1.90
4	3.06	2.46	0.89	1.60	2.76	2.46	0.89	2.02
5	1.60	2.76	0.89	0.89	2.76	0.89	1.60	1.63
6	2.76	1.01	1.60	0.89	1.60	2.76	1.60	1.75
7	1.60	0.17	2.76	0.17	1.01	3.06	2.46	1.60
8	3.06	1.01	1.60	1.60	1.01	0.89	1.01	1.45
9	1.01	1.01	2.76	2.76	1.60	2.46	1.60	1.89
10	3.06	2.76	2.46	0.17	2.46	0.89	1.60	1.91
$\vdots$								$\vdots$
500			...					

Worked out example

Bootstrap estimates:
$\{1.91, 1.54, 1.90, 2.02, 1.63, 1.75, 1.60, 1.45, 1.89, 1.91, \ldots\}$.



Quantile intervals

80% CI: $[1.21, 2.20]$.

95% CI: $[0.92, 2.45]$.

Normal 95% CI: sample mean ± 1.95 sample sd: $[0.97, 2.45]$.

90% CI: sample mean ± 1.65 sample sd: $[1.09, 2.34]$.

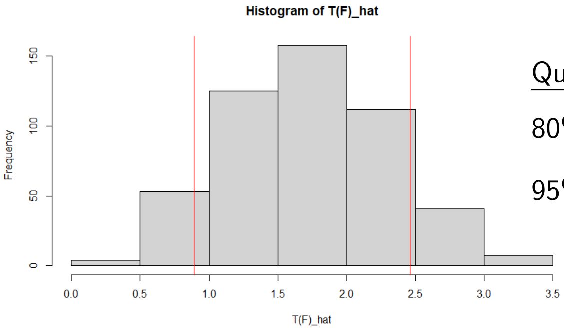
Worked out example number 2

$\mathcal{S} = \{2.6,\ 0.89,\ 1.6,\ 2.76,\ 1.01,\ 3.06,\ 1.17\}$.
$T(F) = \textit{median}$.

b	1	2	3	4	5	6	7	$\widehat{T(F)}$
1	0.17	3.06	3.06	0.89	2.46	2.76	1.01	2.46
2	0.89	1.01	2.46	1.01	0.17	0.17	1.60	1.01
3	0.17	1.60	0.89	0.89	3.06	1.60	2.46	1.60
4	0.17	0.89	1.60	1.01	3.06	1.01	2.46	1.01
5	2.46	2.46	2.46	0.89	1.60	1.60	2.46	2.46
6	0.89	2.46	1.60	0.17	2.46	0.89	3.06	1.60
7	2.46	0.89	0.89	2.76	0.89	0.17	3.06	0.89
8	1.01	1.60	0.17	1.01	1.01	1.01	0.89	1.01
9	0.89	3.06	1.60	0.17	0.17	1.01	2.46	1.01
10	2.76	2.46	1.60	0.17	0.89	0.89	1.01	1.01
$\vdots$								$\vdots$
500			...					

Worked out example number 2

Bootstrap estimates:
$\{2.46,\ 1.01,\ 1.60,\ 1.01,\ 2.46,\ 1.60,\ 0.89,\ 1.01,\ 1.01,\ 1.01,\ \ldots\}$.



Quantile intervals

80% CI: [0.89, 2.46].

95% CI: [0.89, 2.76].

Normal 95% CI: sample mean ± 1.95 sample sd: [0.35, 3.00].

90% CI: sample mean ± 1.65 sample sd: [0.56, 2.79].

Parametric vs non-parametric bootstrap

$\mathcal{S} = \{X_1, X_2, \ldots, X_n\} \sim F(unknown)$.

Bootstrap method is handy when we need to estimate population parameters.
It lets us generate observations from the population as many times as one needs.

Non-parametric Bootstrap

Samples from $\mathcal{S}$ with replacement.
Each sample is of size n.
No distribution assumption.

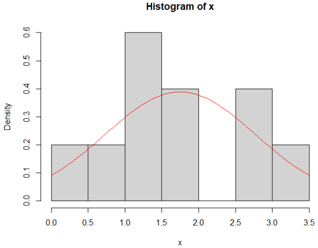
Parametric Bootstrap

Assumes a distribution for F, say, normal $N(\mu, \sigma^2)$ for the data.
Estimates μ and σ from $\mathcal{S}$.
Generates samples from $\widehat{F} = N(\widehat{\mu}, \widehat{\sigma}^2)$.

Parametric bootstrap worked out example

$\mathcal{S} = \{2.6,\ 3.16,\ 0.8,\ 1.19,\ 0.1,\ 1.14,\ 1.73,\ 1.73,\ 2.72,\ 1.05\}$.
$T(F) = median$.

Suppose the true distribution F is normal, $N(\mu, \sigma^2)$.



$\widehat{\mu} = 1.62 =$ sample mean
$\widehat{\sigma} = 0.96 =$ sample sd.

We will generate bootstrap samples from $N(1.62, 0.96^2) = \widehat{F}$.

The procedure after sample generation is the same as the non-parametric bootstrap.

Parametric bootstrap worked out example

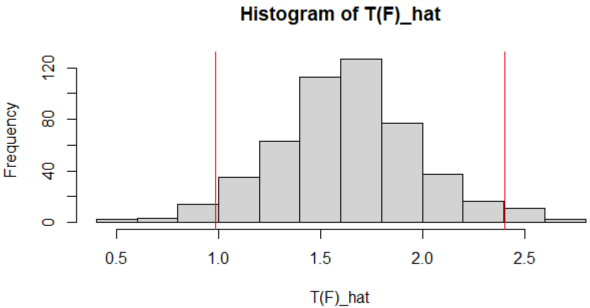
We generated B $= 500$ bootstrap samples from $N(1.62, 0.96^2)$ of size $n = 10$ each.
$T(F) =$ median.

b	1	2	3	4	5	6	7	8	9	10	$\widehat{T(F)}$
1	3.00	2.35	0.23	1.42	1.36	1.25	2.11	1.31	1.00	2.12	1.39
2	2.34	-0.48	1.80	1.37	1.45	1.54	2.45	1.06	1.60	3.70	1.57
3	-0.10	1.91	0.94	3.60	1.70	1.32	1.01	1.34	2.12	1.64	1.49
4	2.04	2.78	1.59	1.47	1.91	3.05	2.49	0.72	1.66	0.39	1.78
5	1.94	2.44	0.50	1.27	2.53	1.01	1.67	0.66	0.53	1.35	1.31
6	3.07	-1.12	1.47	1.79	1.69	2.73	1.08	2.95	1.68	2.11	1.74
7	1.63	1.95	0.46	1.21	1.88	2.34	2.52	0.15	2.70	1.90	1.89
8	2.14	1.20	1.68	3.43	0.38	2.94	1.42	1.32	2.61	2.48	1.91
9	0.47	0.85	2.57	0.88	0.89	3.24	1.71	-0.25	2.75	1.94	1.30
10	2.81	3.62	1.80	2.50	1.23	0.57	1.97	1.03	2.20	1.25	1.88
$\vdots$											$\vdots$
500					...						

Parametric bootstrap worked out example

Bootstrap estimates:

$\{1.39,\ 1.57,\ 1.49,\ 1.78,\ 1.31,\ 1.74,\ 1.89,\ 1.91,\ 1.30,\ 1.88, \ldots\}$.



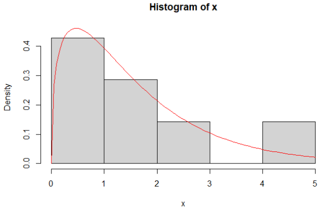
R code

```
> tb <- c(1.39, 1.57, ...)
> quantile(tb, probs = c(0.025, 0.975))
quantile 95% CI: [0.909, 2.32].

> mean(tb) + c(-1.95, 1.95)*sd(tb)
normal 95% CI: [0.910, 2.33].
```

Parametric bootstrap worked out example number 2

$\mathcal{S} = \{2.6,\ 0.89,\ 1.6,\ 0.76,\ 1.01,\ 4.06,\ 0.17\}$.
$T(F) = mean$.

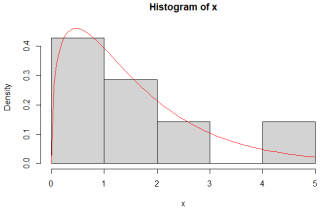


$$\widehat{\alpha} = 1.41 = \frac{\textit{sample mean}^2}{\textit{sample variance}}$$
$$\widehat{\beta} = 0.89 = \frac{\textit{sample mean}}{\textit{sample variance}}.$$

We will generate bootstrap samples from $Gamma(1.41, 0.89) = \widehat{F}$.

Parametric bootstrap worked out example number 2

$\mathcal{S} = \{2.6,\ 0.89,\ 1.6,\ 0.76,\ 1.01,\ 4.06,\ 0.17\}.$
$T(F) = mean.$



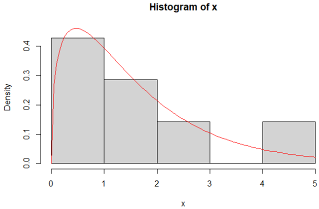
$$\widehat{\alpha} = 1.41 = \frac{sample\ mean^2}{sample\ variance}$$
$$\widehat{\beta} = 0.89 = \frac{sample\ mean}{sample\ variance}.$$

We will generate bootstrap samples from $Gamma(1.41, 0.89) = \widehat{F}$.

Parametric bootstrap worked out example number 2

$\mathcal{S} = \{2.6,\ 0.89,\ 1.6,\ 0.76,\ 1.01,\ 4.06,\ 0.17\}$.
$T(F) = mean$.

Suppose the true distribution F is gamma, $Gamma(\alpha, \beta)$.



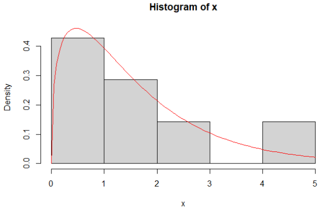
$$\widehat{\alpha} = 1.41 = \frac{sample\ mean^2}{sample\ variance}$$
$$\widehat{\beta} = 0.89 = \frac{sample\ mean}{sample\ variance}.$$

We will generate bootstrap samples from $Gamma(1.41, 0.89) = \widehat{F}$.

Parametric bootstrap worked out example number 2

$\mathcal{S} = \{2.6,\ 0.89,\ 1.6,\ 0.76,\ 1.01,\ 4.06,\ 0.17\}$.
$T(F) = mean$.

Suppose the true distribution F is gamma, $Gamma(\alpha, \beta)$.



$$\widehat{\alpha} = 1.41 = \frac{\textit{sample mean}^2}{\textit{sample variance}}$$
$$\widehat{\beta} = 0.89 = \frac{\textit{sample mean}}{\textit{sample variance}}.$$

We will generate bootstrap samples from $Gamma(1.41, 0.89) = \widehat{F}$.

The procedure after sample generation is the same as the non-parametric bootstrap.

Precaution: The underlying distribution F is different from the bootstrap estimates distribution.

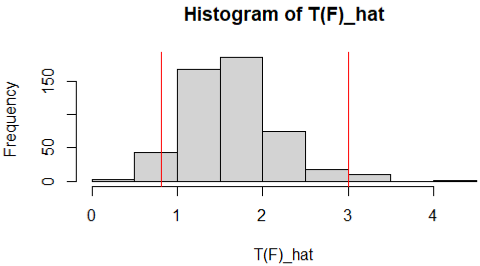
Parametric bootstrap worked out example 2

We generated B = 500 bootstrap samples from
$Gamma(1.41, 0.89)$ of size $n = 7$ each.
$T(F)$ = mean.

b	1	2	3	4	5	6	7	$\widehat{T(F)}$
1	1.44	0.48	1.15	2.69	1.93	0.46	1.11	1.32
2	0.14	3.53	0.43	1.82	6.71	0.24	0.20	1.87
3	0.44	0.77	0.31	1.58	3.76	0.85	1.25	1.28
4	2.95	0.41	4.27	0.03	4.18	1.21	0.26	1.90
5	0.61	2.06	5.29	1.00	2.23	1.56	2.42	2.17
6	1.39	0.22	3.56	0.77	0.29	1.64	0.38	1.18
7	0.27	1.47	1.57	0.35	1.47	0.26	0.17	0.79
8	0.29	1.47	2.19	0.13	2.25	1.32	1.77	1.34
9	3.24	1.67	1.74	0.05	1.60	1.95	0.68	1.56
10	1.97	0.27	0.71	2.86	5.20	1.06	2.06	2.02
⋮								⋮
500				...				

Parametric bootstrap worked out example 2

Bootstrap estimates:
$\{1.32,\ 1.87,\ 1.28,\ 1.90,\ 2.17,\ 1.18,\ 0.79,\ 1.34,\ 1.56,\ 2.02, \ldots\}$.



The distribution of the bootstrap estimates is roughly normal because of central limit theorem given n is sufficiently large.

R code

```
> tb <- c(1.32, 1.87, ...)
> quantile(tb, probs = c(0.025, 0.975))
quantile 95% CI: [0.81, 3.00].

> mean(tb) + c(-1.95, 1.95)*sd(tb)
normal 95% CI: [0.63, 2.67].
```

Advantages of non-parametric and parametric bootstrap

Non-parametric

1. No distribution assumption.

Disadvantage: Repetitive values for small samples.

Parametric

1. Can be used even for small samples.
2. The bootstrap samples are less redundant compared to non-parametric bootstrap.

Disadvantage: Needs model assumption.

Summary and points to remember

We have seen the methods and examples of non-parametric and parametric bootstrap for simple problems to study population parameters $T(F)$.

- Original data comes from distribution F whereas, bootstrap sample comes from an approximation of F, i.e. $\widehat{F}$.
- Bootstrap sample size is the same as the original sample size.
- Bootstrap sample estimates are roughly normal irrespective of F by central limit theorem.
- We can obtain confidence interval of $T(F)$ from bootstrap estimates by either quantile method or by normal confidence interval.

Breakout rooms

Calculate 95% confidence interval for the median of the toy data in `originalsample.txt`.

Generate B bootstrap samples and get the confidence intervals using quantiles and normal for $B = 50, 1000$.

Open `Breakout_activity_1.pdf`

Optional: **Bootstrap in linear regression**

Bootstrap use in model fitting

We can use bootstrap to obtain a more robust model estimate in linear regression.

Data: $\{(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)\}$.

Simple regression: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

There are two approaches for this

1. random x
   Sample from $\{(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)\}$ with replacement.
2. fixed x
   Sample from $\{\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n\}$ via non-parametric or parametric bootstrap method.

Random x bootstrap in linear regression

Original data: $\begin{bmatrix} Y_1 & x_1 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ Y_4 & x_4 \\ Y_5 & x_5 \\ Y_6 & x_6 \end{bmatrix}$. Fitted model $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

Draw B $=$ 5 bootstrap samples from the data:

$$\begin{bmatrix} Y_1 & x_1 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ Y_1 & x_1 \\ Y_3 & x_3 \\ Y_3 & x_3 \end{bmatrix} \quad \begin{bmatrix} Y_2 & x_2 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ Y_3 & x_3 \\ Y_6 & x_6 \\ Y_6 & x_6 \end{bmatrix} \quad \begin{bmatrix} Y_4 & x_4 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ Y_1 & x_1 \\ Y_5 & x_5 \\ Y_6 & x_6 \end{bmatrix} \quad \begin{bmatrix} Y_6 & x_6 \\ Y_3 & x_3 \\ Y_3 & x_3 \\ Y_4 & x_4 \\ Y_4 & x_4 \\ Y_6 & x_6 \end{bmatrix} \quad \begin{bmatrix} Y_2 & x_2 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ Y_3 & x_3 \\ Y_2 & x_2 \\ Y_6 & x_6 \end{bmatrix}$$

Fit a linear model with each data sample to get B $=$ 5 fitted models $\widehat{Y}^{(b)} = \widehat{\beta}_0^{(b)} + \widehat{\beta}_1^{(b)} x, \ b = 1, \ldots, 5.$

Random x bootstrap in linear regression

Final fitted model is the average of all the $B = 5$ fitted models

$$\widehat{Y} = \frac{1}{B}\sum_b \widehat{\beta}_0^{(b)} + \frac{1}{B}\sum_b \widehat{\beta}_1^{(b)} x.$$

This method is equivalent to the non-parametric bootstrap but for data pairs (Y_i, x_i). Also, applicable for $(Y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$.

To get confidence interval or variance of β_0 and β_1, use the bootstrap estimates

$$\{\widehat{\beta}_0^{(b)} : b = 1, \ldots, B\} \quad \text{and} \quad \{\widehat{\beta}_1^{(b)} : b = 1, \ldots, B\}.$$

Calculate either the quantile interval or the nominal confidence interval.

Fixed x bootstrap in linear regression

Original data: $\begin{bmatrix} Y_1 & x_1 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ \vdots & \vdots \\ Y_n & x_n \end{bmatrix}$. Fitted model $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

Obtain the fitted residuals $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i,\ i = 1, \ldots, n$.

The sampling procedure to get new training data is performed on

$$\mathcal{S} = \{\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n\}.$$

Non-parametric Draw a sample $\{\widehat{\epsilon}_1^{(b)}, \ldots, \widehat{\epsilon}_n^{(b)}\}$ from $\mathcal{S}$ with replacement.

Parametric Draw a sample $\{\widehat{\epsilon}_1^{(b)}, \ldots, \widehat{\epsilon}_n^{(b)}\}$ from $N(mean(\widehat{\epsilon}), \widehat{\sigma}^2)$.

Fixed x bootstrap in linear regression

Original data: $\begin{bmatrix} Y_1 & x_1 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ \vdots & \vdots \\ Y_n & x_n \end{bmatrix}$. Fitted model $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

Fitted residuals are $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i, \ i = 1, \ldots, n$.

Bootstrap sample for $b = 1, \ldots, B$ of residuals is $\{\widehat{\epsilon}_1^{(b)}, \ldots, \widehat{\epsilon}_n^{(b)}\}$.

Bootstrap training data for $b = 1, \ldots, B$ is

$$\begin{bmatrix} \widehat{Y}_1 + \widehat{\epsilon}_1^{(b)} & x_1 \\ \widehat{Y}_2 + \widehat{\epsilon}_2^{(b)} & x_2 \\ \widehat{Y}_3 + \widehat{\epsilon}_3^{(b)} & x_3 \\ \vdots & \vdots \\ \widehat{Y}_n + \widehat{\epsilon}_n^{(b)} & x_n \end{bmatrix}.$$

Regress $Y^* = \widehat{Y} + \widehat{\epsilon}^{(b)}$ on x to get $\widehat{Y^*} = \widehat{\beta}_0^{(b)} + \widehat{\beta}_1^{(b)} x$.

$$\widehat{Y} = \frac{1}{B} \sum_b \widehat{\beta}_0^{(b)} + \frac{1}{B} \sum_b \widehat{\beta}_1^{(b)} x.$$

Summarizing bootstrap in linear regression

Original data: $\begin{bmatrix} Y_1 & x_1 \\ Y_2 & x_2 \\ Y_3 & x_3 \\ \vdots & \vdots \\ Y_n & x_n \end{bmatrix}$. Fitted model $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

Random $x \longrightarrow$ Draw $(Y^{(b)}, x^{(b)})$ from original data with replacement

Fixed x

- Non-parametric $\longrightarrow$ Draw from $\widehat{\epsilon}_i$ with replacement
  New data is $Y + \widehat{\epsilon}^{(b)}$, x
- Parametric $\longrightarrow$ Draw from $N\left(mean(\widehat{\epsilon}), \widehat{\sigma}^2\right)$
  New data is $Y + \widehat{\epsilon}^{(b)}$, x