

**Predictive Analytics - Regression and Classification**  
**Prof. Sourish Das**  
**Department of Mathematics**  
**Chennai Mathematical Institute**




**Lecture - 32**  
**Bootstrap Regression**

(Refer Slide Time: 00:26)

**Bootstrap Statistics**

Data:  $y = \{y_1, y_2, \dots, y_n\}$   $\xrightarrow{\text{random sample}}$   $[y_1^*, y_2^*, \dots, y_n^*]$   $\xrightarrow{\text{resample}}$

- ▶ Bootstrap statistics is an algorithmic strategy, which typically resort to SRSWR scheme
- ▶ It falls under the broader 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.



Hello all welcome back to the part C of lecture 9. In this part we are going to talk about the Bootstrap Regression and bootstrap statistics is an algorithmic strategy, which typically resort to simple random sample with replacement strategy or sometimes without replacement, but mostly it resort to with replacement strategy. Now, it falls under the broader class of resampling strategy; that means, you if you have a sample say if you have a sample  $y$ .

Which have some values  $y_1, y_2, \dots, y_n$ . What we will do? We will draw this is main data ok. This you can consider as your main data and this is your main data and what we will do? We will draw random samples from this data ok. So, we will draw random samples from

this data. Now, this itself is this data itself is a sample from the population. Now, we are doing random sample this from this sample.

So, that is why this new sample maybe  $y_1$ ,  $y_2$ , ...,  $y_n$ ; this sample will be called resample. So, resample and this is a broad class of broad class of algorithms are there and this strategies are called resampling strategy or resampling algorithms. This idea of bootstrap was first introduced by Brad Efron Professor Brad Efron in 1979 in his Analysis of Statistics paper.

The idea though apparently simple, it is a very simple idea, but it revolutionized the statistics by its ability to replace analytical derivation by brute force computing. So, this was I think was very timely with respect to an advancement of computer, personal computer etcetera and this pretty much revolutionized the idea.

(Refer Slide Time: 02:53)

### Bootstrap Statistics

$F(\cdot)$  could any prob dist?!


- ▶ Suppose  $\{Y_1, Y_2, \dots, Y_n\}$  are iid observations with cdf  $F(\cdot)$  and  $T_n = T_n(Y_1, Y_2, \dots, Y_n)$  is a statistic which estimates a parameter  $\theta$ .
- $\theta$  is some parameter of  $F(\cdot)$ .
- ▶ 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  $F_n(\cdot)$ .


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

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

$\mathbb{I}(A) = 1$  if  $A$  is true  
 $= 0$  otherwise

- ▶ Bootstrapping based on  $F_n(\cdot)$  is called nonparametric





First, I will present the idea of bootstrap statistics with respect to one univariate sample or univariate variable and then I will describe once you get the idea of bootstrap statistics what I will do? I will represent I will present the idea of bootstrap statistics in predictive model or regression context.

So, suppose  $Y_1, Y_2, Y_n$  are IID observation from any probability distribution. So, we are calling it say  $F$ . Now,  $F$  could be normal,  $F$  could be Gaussian anything. So,  $F$  could be any probability distribution, any probability distribution ok. So, in and  $T_n$ ,  $T_n$  will be some function of the  $\theta$ ;  $T_n$  will be some function of the  $\theta$   $Y_1, Y_2, Y_n$ .

This is a statistics which estimate the parameter  $\theta$  ok. So,  $\theta$  is some parameter of  $F$  is some parameter of  $F$ , parameter of  $F$ ,  $F$  is the probability distribution completely unknown and what you are using? You are using  $T_n$ , the sampling which is a function of the sample. It tries to estimate,  $T_n$  tries to estimate the  $\theta$ .

Now, as usual the sampling distribution of  $T_n$  will depend on  $F$  sampling distribution will depend on the population distribution  $F$  is the population distribution,  $F$  is the population distribution. So, the bootstrap idea is simple in simplest form is to estimate; so,  $F$  is unknown the population distribution is unknown.

So, what we will do? We will estimate the unknown population distribution by the empirical probability distribution. What is empirical probability distribution? Very simple; empirical probability distribution only  $x$  defined as something like this;  $\frac{1}{n} \sum_{i=1}^n I_{x_i}$  ok. So, what is indicator function?

Indicator function if condition  $A$  is satisfied then I will call 1 or 0 if  $A$  is true 0 otherwise. So, that is how it is being defined. It is a simple function you can define it; it is very easy that just like you know you can write a for loop and define the function in any program it just like 3 lines of code.

So, the empirical CDF is and here is the big result. The big result is tells you the theoretical result that tells you that empirical CDF of  $F_n$  CDF that you know this function is a non-parametric maximum MLE estimate of the CDF unknown CDF.

So, if you have a unknown probability distribution the best case that about that unknown distribution you can make is through this particular distribution. Empirical CDF  $F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$ . This indicator this empirical CDF you can code it in like you know single pretty much you know in 3 to 4 lines that is it 3 4 lines of Python code or R code.

(Refer Slide Time: 07:42)

4 of 20 sult The empirical cdf  $F_n(\cdot)$  is the non-parametric MLE of cdf  $F(\cdot)$ .

$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$

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

**Bootstrap Statistics**

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

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

So, the bootstrap pin based on this empirical CDF is called non-parametric bootstrap. Now, bootstrap statistics is suppose,  $y_1, Y_2, Y_n$  are iid observations with cdf  $F$  and  $T_n$  is  $T_n, Y_1, Y_2, Y_n$  is a statistics with estimates the parameter  $\theta$  ok. So, that since the empirical

CDF is the non parameter non-parametric MLE of CDF. Now, what we can do? We can draw sample from  $F_n(x)$ . This empirical CDF we can we know how to draw sample from empirical CDF.

(Refer Slide Time: 08:45)

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  $\{Y_1, Y_2, \dots, Y_n\}$
- ▶ That is draw resamples from  $\{Y_1, Y_2, \dots, Y_n\}$
- ▶ Hence we can draw as many times as we want.

$\{Y_1^*, Y_2^*, \dots, Y_n^*\}$

cmj

NPTEL

Bootstrap Framework

- ▶  $Y_n = \{Y_1, Y_2, \dots, Y_n\}$  are iid random samples from  $F(\cdot)$ .

This is simply now turns out the drawing empirical from empirical CDF is same as drawing sample iid samples from your data  $Y_1, Y_2, Y_n$  that is it. So, that is why the resampling comes in. So, you draw sample from if you draw resample that is if you resample from  $Y_1, Y_2, Y_n$ .



Suppose that comes  $Y_1^*, Y_2^*, \dots$  sorry for that, let me just write it carefully;  $Y_1^*, Y_2^*, \dots, Y_n^*$ . Suppose this is the resample data from  $Y_1, Y_2, Y_n$ . Now, these resamples are basically same as drawing sample from empirical CDF which is maximum likelihood estimate for unknown probability distributions. And hence we can draw

as many times we can repeat this resampling as many times as we want. So, that is the main idea of bootstrap statistics.

(Refer Slide Time: 10:06)

### Bootstrap Framework

- ✓  $\mathbf{Y}_n = \{Y_1, Y_2, \dots, Y_n\}$  are iid random samples from  $F(\cdot)$
- ✓  $\hat{T}_n = T_n(\mathbf{Y}_n)$  is a statistic for parameter  $\theta$ 
  - ▶ 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.,  $\text{Var}(T_n)$  and confidence interval of  $T_n$ , i.e.,  $CI(T_n)$ .
  - ▶ Resample  $\mathbf{Y}_{nb}^* = \{Y_1^*, Y_2^*, \dots, Y_n^*\}_b$  from  $\mathbf{Y}_n$  using SRSWR scheme;  $b = 1, 2, \dots, B$

So, let me explain you the bootstrap framework now. What is the framework? So, you have these you are starting with this data set that this is the data set that you have and you are assuming these are iid random samples from a unknown distribution F of x or F of y some unknown distribution F completely unknown. You have no idea it could be normal comma normal you have no idea.

Now, you can also do you can estimate the  $T_n$  with some method using some function you can estimate the  $T_n$  which is a statistic for estimating parameter theta. Estimation is not a issue. What is in problem is you do not know what is F. So, you do not know what is the sampling distribution of  $T_n$ .

So, you do not know what is the variance of  $T_n$  is unknown variance of  $T_n$  is unknown. Confidence interval of  $T_n$  is unknown to you. So, you cannot do you cannot compute the margin of error you cannot do any statistical inference. So, what is the approach? Then you resample  $Y_{nb}$  star from  $Y_n$  from here from this data set from this data set you just from this data set you just resample.

These are your resamples and you resample many many many many times as many times as possible. So, this capital  $B$  is you decide, it could be 1000 times, it could be 10,000 times, it could be 100,000 times, it could be 1 million times. You decide how many times you depending on your value for computational ability you decide.

(Refer Slide Time: 12:10)

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  $Y_{nb}^* = \{Y_1^*, Y_2^*, \dots, Y_n^*\}_b$  from  $Y_n$  using SRSWR scheme;  $b = 1, 2, \dots, B$ .
- For each resample  $b$ , we can compute  $T_{nb}^*$ ,  $b = 1, 2, \dots, B$ .

Bootstrap Framework

histogram

$T_{n1}^* \quad T_{n2}^* \quad T_{n3}^* \quad \dots \quad T_{nB}^*$

$Y_n = \{Y_1, Y_2, \dots, Y_n\}$  are iid random samples from  $F(\cdot)$ .

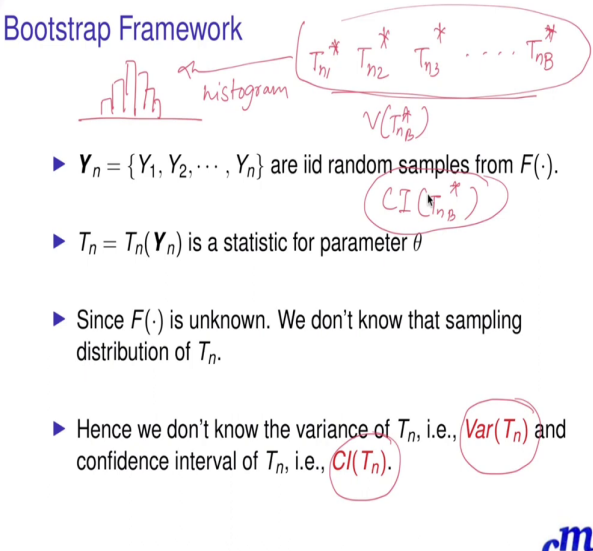
Once you compute, once you resample once you re sample this  $Y_{nb}$  for this  $Y_{nb}$  you can calculate again the statistics  $T_{nb}$ . So, for each resample  $b$  we can compute this  $T_{nb}$  star and

once I get all these  $T_{nB}$  stars. So, why now I have this resample estimate of  $T_{n1}$ ,  $T_{n2}$  star  $T_{n3}$  star all these estimates  $T_{nB}$  star.



Now, based on these data I can have a sampling estimates of the sampling distributions. Some I can draw the histogram ok. I can draw the histogram and I can get a sense of what is the how the histogram looks like ok.

(Refer Slide Time: 13:12)

### Bootstrap Framework



- ▶  $Y_n = \{Y_1, Y_2, \dots, Y_n\}$  are iid random samples from  $F(\cdot)$ .
- ▶  $T_n = T_n(Y_n)$  is a statistic for parameter  $\theta$
- ▶ 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.,  $CI(T_n)$ .

And based on this histogram I can calculate what is the or these values we can calculate what is the variance of  $T_n$  star  $T_{nB}$  star and what is the confidence interval of  $T_{nB}$  stars. So, remember that we do not what we do not do is variance of  $T_n$  and confidence interval of  $T_n$ , but using  $T_{nB}$  star I can calculate variance of  $T_{nB}$  star and confidence interval of  $T_{nB}$  star.



(Refer Slide Time: 13:45)

### Bootstrap Framework




- ▶ Resample  $\mathbf{Y}_{nb}^* = \{Y_1^*, Y_2^*, \dots, Y_n^*\}_b$  from  $\mathbf{Y}_n$  using SRSWR scheme;  $b = 1, 2, \dots, B$
- ▶ For each resample  $b$ , we can compute  $T_{nb}^*$ ;  $b = 1, 2, \dots, B$   
 $T_{n1}^* \quad T_{n2}^* \quad \dots \quad T_{nB}^*$
- ▶ We can compute:

$$\bar{T}_n^B = \frac{1}{B} \sum_{b=1}^B T_{nb}^* \quad \text{Var}(T_n)^B = \frac{1}{B} \sum_{b=1}^B (T_{nb}^* - \bar{T}_n^B)^2$$

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

$$T_n + G_B^{-1}(1 - \alpha/2) \sqrt{\text{Var}(T_n)^B}\}$$

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

Now, so let me repeat the bootstrap framework again. So, from the main data we resample the data and for each resample we compute the statistics. And based on statistics, so they basically  $T_{n1}^*$ ,  $T_{n2}^*$  for each resamples we compute  $T_n$  capital B star, capital B you decide user decide it could be 1000, 10,000 whatever. Now, based on B's computation you can compute take the average of these guys. You can calculate the variance of these guys, you can calculate the confidence interval of these guys.

(Refer Slide Time: 14:38)

8 of 20


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

cm

NPTEL

### Bootstrap Framework

► Due to SLLN, one can show, as  $B \rightarrow \infty$

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


Now, from you can also define empirical distribution  $G_B$  this guy is scaled  $G_B$ . So, all these things you can do. Now due to strong law of large number as capital  $B$  goes to infinity; remember, the capital  $B$  is user different. So, you can make it as much large as possible ok.

(Refer Slide Time: 15:01)

### Bootstrap Framework

► Due to SLLN, one can show, as  $B \rightarrow \infty$ ,

$\bar{T}_n^B \rightarrow T_n$  almost surely;

$\text{Var}(T_n)^B \rightarrow \text{Var}(T_n)$  almost surely

$CI(T_n)^B \rightarrow CI(T_n)$  almost surely,



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

$F(\cdot)$  unknown

$\hookrightarrow F_{T_n}(\cdot)$  "

$\hookrightarrow \text{Var}(T_n)$  "

$\hookrightarrow CI(T_n)$  "

One can show  $\bar{T}_n^B$  converges to  $T_n$  almost surely. The variance of  $\bar{T}_n^B$  converges to variance of  $T_n$  almost surely. Confidence of confidence interval converges to confidence interval of like resampled  $\bar{T}_n^B$  is converges to confidence interval almost surely. And sampling distribution of bootstrap sampling distribution converges to very  $F$  of  $T_n$ .

Now, what is happening? What we do not know? Because  $F$  is unknown because the population distribution is unknown,  $F$  is unknown true population distribution is unknown. So, because of that  $F_{T_n}$  is unknown. Because  $F_{T_n}$  is unknown because  $F_{T_n}$  is unknown variance of  $T_n$  is unknown because if variance of  $T_n$  is unknown you cannot compute the confidence interval of  $T_n$ . So, that is the main problem that we are facing here.



But what bootstrap statistics is saying that you do not have to worry all you have to ensure you have enough computational capacity, push the B to infinity and variance of  $T_n$  we will converge to variance of T, variance of bootstrap  $T_n$  will converge to the variance of  $T_n$ .

So, that means, basically variance of bootstrap  $T_n$  will be same as variance of  $T_n$ . Similarly, confidence interval the bootstrap, confidence interval will be same as bootstrap actual confidence interval of the  $T_n$ . So, this is the strong strength of the work.

(Refer Slide Time: 16:51)

### Bootstrap Regression

- ▶ Consider the model
 
$$\mathbf{y}_n = \mathbf{X}_{n \times p} \beta_p + \epsilon_n,$$
- where  $\mathbb{E}(\epsilon) = 0$ ,  $\text{Var}(\epsilon) = \sigma^2 \mathbf{I}_n$  and  $\epsilon \stackrel{iid}{\sim} F(\cdot)$ ,  $F(\cdot)$  is unknown cdf
- ▶ OLS estimator:  $\hat{\beta}_n = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ ; and  $\text{Var}(\hat{\beta}_n) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ .
- ▶ Residuals:  $\hat{\epsilon} = \mathbf{y} - \mathbf{X} \hat{\beta}_n$  or  $\hat{\epsilon}_i = y_i - \mathbf{x}_i^T \hat{\beta}_n$ ,  $i = 1, 2, \dots, n$ .

Now, I am this is the idea of the bootstrap statistics. Now, I am going to explain you bootstrap statistics in context of regression. So, now we are going to talk about bootstrap regression. So, let us consider the model  $y_n, X_n \times \beta$  plus epsilon. Expectation of epsilon is 0, variance of epsilon in sigma square  $I_n$ .

So, homoscedasticity is still I am holding. But what is what I am now giving up is normality. So, most of the time what we are seeing that at least in the capital asset pricing model what we have found that homoscedasticity was ok even the randomness was ok.

But what was not holding good was residual was definitely not normal. Since the residual was not normal. So, now what we are seeing what we are doing we are saying that ok residuals are coming from iid distribution, but we do not know what is the distribution. So, we are seeing  $f$  is unknown CDF. Since  $F$  is unknown CDF, we cannot do any inference on  $\beta$ .




So, we can estimate the OLS estimator. OLS estimator is all you do is  $X^T X^{-1} X^T y$  that not a big deal. You can even calculate variance of  $\beta_n$   $\hat{\sigma}^2 X^T X^{-1} y$ . But you cannot do the confidence interval. You cannot calculate the confidence interval. So, that is where we are getting stuck.

So, what we can do is we can look into the residual, what is residual? Residuals are  $\epsilon$  equal to  $y$  minus  $X \beta_n$   $\hat{\sigma}$ . We can either equal to this and here or error equal to  $e_i$  equal to  $y_i$  minus  $x_i^T \beta_n$   $\hat{\sigma}$ . It should be not  $\epsilon$ , it should be  $e_i$ . Because it is observed we are using  $X \beta_n$   $\hat{\sigma}$  here.

(Refer Slide Time: 19:09)

### Residual Bootstrap Regression

- ▶ Suppose  $F_n(\cdot)$  is the empirical cdf of  $\epsilon$
- ▶  $\epsilon_b^* \stackrel{iid}{\sim} F_n$  (i.e.,  $\epsilon_b^*$  is resampled from  $\epsilon$  using SRSWR),  
 $b = 1, 2, \dots, B = 10000$
- ▶ Calculate:
 
$$\mathbf{y}_b^* = \mathbf{X}\hat{\beta}_n + \epsilon_b^*$$
- ▶ Estimate resample coefficients  $\hat{\beta}_{n,b}^*$  as
 
$$\begin{aligned} \hat{\beta}_{n,b}^* &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}_b^* \\ &= \hat{\beta}_n + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon_b^* \\ \mathbb{E}(\hat{\beta}_{n,b}^*) &= \hat{\beta}_n \end{aligned}$$
- ▶ Bootstrap Estimate:  $\bar{\beta}_B = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b^*$
- ▶ Bootstrap variance:  $\text{Var}(\bar{\beta}_B) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_b^* - \bar{\beta}_B)^2$

Now, what we can. So, we can have these error observed error. Now, suppose  $F_n$  is the empirical CDF of these error ok small  $\epsilon$ . So, now what we are saying that alright these  $\epsilon_i$  is  $\epsilon_b^*$  follow iid  $F_n$  then  $\epsilon_b^*$  is resampled from  $\epsilon$  using SRS with SRSWR where  $b$  equal to 1 to  $B$ ,  $B$  capital  $B$  could be anything any big number 10,000 or any 100,000 maybe whatever.

Now, you calculate all you have to do just you know what is  $\mathbf{X}\hat{\beta}_n$  just add epsilon  $\epsilon_b^*$ . Then you get a new response  $\mathbf{y}_b^*$ . And once you get the  $\mathbf{y}_b^*$  all you have to do you get estimate the resampled coefficient you can estimate as  $\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T \mathbf{y}_b^*$  with the new response; new bootstrap response.

But turns out you can write it as  $\hat{\beta}_n$  which you already know plus  $\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T \epsilon_b^*$  ok. Now,  $\mathbf{X}^T \mathbf{X}^{-1} \mathbf{X}^T$  you already know

right. So,  $e_b$  star all you have to do is multiply and that will give you the  $\beta_n$  star hat. And one can show that expected value of  $\beta_n$  hat star bootstrap is  $\beta_n$  hat. One can show that it is not a very difficult thing to show. So, what is my bootstrap estimate? Bootstrap estimate is simply you take the average of all these  $B$  sample estimates and calculate the variance of all these resampled estimates ok.

(Refer Slide Time: 21:14)

10 of 20

- ▶  $\epsilon_b^* \sim F_n$  (i.e.,  $\epsilon_b^*$  is resampled from  $\epsilon$  using SRSWR),  $b = 1, 2, \dots, B = 10000$
- ▶ Calculate:
 
$$y_b^* = X\hat{\beta}_n + \epsilon_b^*$$
- ▶ Estimate resample coefficients  $\hat{\beta}_{n,b}^*$  as
 
$$\hat{\beta}_{n,b}^* = (X^T X)^{-1} X^T y_b^*$$

$$= \hat{\beta}_n + (X^T X)^{-1} X^T \epsilon_b^*$$

$$\mathbb{E}(\hat{\beta}_{n,b}^*) = \hat{\beta}_n$$
- ▶ Bootstrap Estimate:  $\bar{\beta}_B = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b^*$
- ▶ Bootstrap variance:  $\text{var}(\bar{\beta}_B) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_b^* - \bar{\beta}_B)^2$

NPTEL

cmi

Paired Bootstrap Regression

So, this is my bootstrap estimates. Now, this is and based on these bootstrap estimate and bootstrap variance you can do the simple covariate statistical inference. Now, there is no problem because you are not even making any assumptions. Now, there is a another approach this method was called residual bootstrap regression ok.

This method was called residual bootstrap regression because you are doing it resampling the residuals. You are resampling the residuals ok.

(Refer Slide Time: 22:14)

▶ Consider the model




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

where  $\mathbb{E}(\boldsymbol{\epsilon}) = 0$ ,  $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$ , and  $\boldsymbol{\epsilon} \stackrel{iid}{\sim} F(\cdot)$ ,  $F(\cdot)$  is unknown cdf

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

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

*Homoscedasticity is okay*





(Refer Slide Time: 22:29)



11 of 20 ▶ Bootstrap variance:  $\text{Var}(\bar{\beta}_B) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_b^* - \bar{\beta}_B)^2$

**Paired Bootstrap Regression**

- ▶ Consider the model
$$\mathbf{y}_n = \mathbf{X}_{n \times p} \beta_p + \epsilon_n,$$

where  $\mathbb{E}(\epsilon) = 0$ ,  $\text{Var}(\epsilon) = \Sigma$ , and  $(y_i, \mathbf{x}_i) \stackrel{iid}{\sim} F(\cdot)$ ,  $F(\cdot)$  is unknown cdf

- ▶ Suppose  $\{(y_i^*, \mathbf{x}_i^*), i = 1, 2, \dots, n\}_b = \mathcal{D}_b$  are iid samples from empirical  $F_n(\dots)$ , where  $b = 1, 2, \dots, B$
- ▶ The estimates of  $\beta$  from  $b^{\text{th}}$  resample:
$$\hat{\beta}_b^* = (\mathbf{X}_b^{*T} \mathbf{X}_b^*)^{-1} \mathbf{X}_b^{*T} \mathbf{y}_b^*$$



(Refer Slide Time: 22:30)


11 of 20

- ▶ Suppose  $\{(y_i^*, \mathbf{x}_i^*), i = 1, 2, \dots, n\}_b = \mathcal{D}_b$  are iid samples from empirical  $F_n(\dots)$ , where  $b = 1, 2, \dots, B$
- ▶ The estimates of  $\beta$  from  $b^{\text{th}}$  resample:
$$\hat{\beta}_b^* = (\mathbf{X}_b^{*T} \mathbf{X}_b^*)^{-1} \mathbf{X}_b^{*T} \mathbf{y}_b^*$$
- ▶ Bootstrap Estimate:  $\bar{\beta}_B = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b^*$
- ▶ Bootstrap variance:  $\text{Var}(\bar{\beta}_B) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}_b^* - \bar{\beta}_B)^2$

cm


NPTEL

### Bootstrap Regression





(Refer Slide Time: 22:33)

Bootstrap Regression



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



Now, suppose, but you are still you are assuming that your homoscedasticity holds good. Homoscedasticity is ok. Your data still preserve the homoscedasticity, but we have seen previously a cases where homoscedasticity even not working. Now, there were cases where we have seen there were studies where we have seen that homoscedasticity was not working.

In that case, you cannot use residual bootstrap regression. Because residual bootstrap regression still assume that your data is has homoscedasticity. But if the homoscedasticity does not work then you better go for heterosceda paired bootstrap model. In the paired bootstrap model what you do is very simple.

So, you are assuming  $y$  same model  $y$  equal to  $X$  beta plus epsilon, but now you are assuming variance of epsilon could be any sigma. It could be any sigma and  $y_i$   $x_i$  star are following  $F$  some unknown CDF. Now, from the  $D$  b from the data size you just draw resample the  $y_i$


star  $x_i$  star as a paired as a pair you resample and you resample many many many many times and do it for each resample you have to calculate these OLS estimates.

Remember that in this case what will happen is the OLS estimates are not necessarily you have to calculate this inverse all the time. In the residual in each the sample you have to calculate this inverse, but in residual bootstrap you do not have to calculate this inverse because, you do it once and you are done all you have to do you have to just draw the residual sample and add this to the residual sample that is it ok.

So, in this case, but you have to do this and then the simple bootstrap estimate and confidence estimates and variance and confidence interval can be done in the regular way. So, if the residuals are heteroscedastic then paired bootstrap is still consistent estimated and you better use paired bootstrap and try to avoid the residual bootstrap. However, in case of heteroscedastic residual the residual bootstrap is not a consistent estimator.

(Refer Slide Time: 25:30)


Paired Bootstrap Regression



	Estimate	Std. Error	t value	Pr(> t )
alpha	-0.0009	0.0010	-0.8790	0.3809
beta	1.0150	0.1157	8.7748	0.0000

Paired Bootstrap Estimates of alpha and beta

	Estimate	Std. Error	2.5%	97.5%
alpha	-0.0009	0.0010	-0.0030	0.0009
beta	1.0137	0.0915	0.8415	1.2004



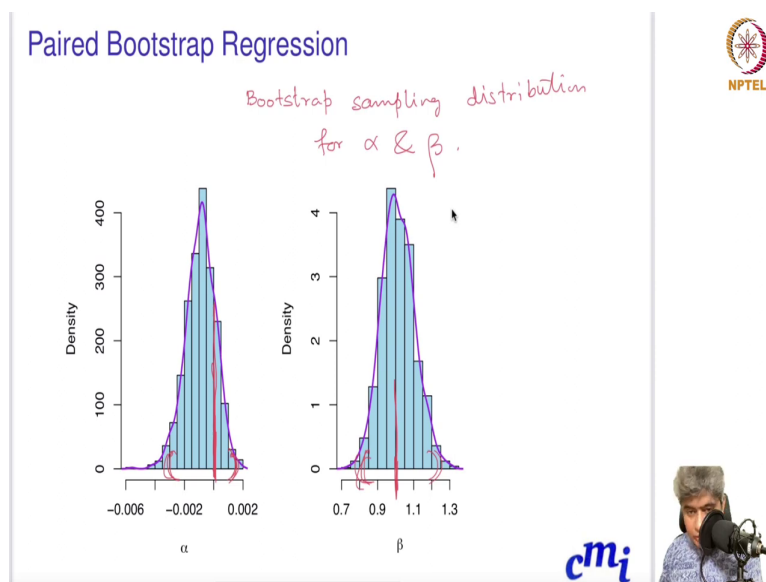
cm

So, be careful about this. So, here is some example that what we found that in the OLS estimate with the you know simple z t value based regression analysis which assumes the residuals follow normal distribution in paired bootstrap case. What we found is that the estimates bootstrap estimates is actually converging to the OLS estimates.

So, in that way estimates is not a problem and almost close to the both alpha and beta very close to the OLS estimates. Standard error can also be somewhat similar, but little less in case in this case. However, and we can compute the confidence interval. Perhaps this is the correct confidence interval given for the OLS estimates given; that we are assuming that the original distribution could be anything the residual distribution could be anything.

Whereas this was assuming normal and we found that the normality is not valid assumptions and in the paired bootstrap you do not even need the assumption of the homoscedasticity.

(Refer Slide Time: 26:58)



So, this is an interesting thing. So, here is the histogram of the paired bootstrap regression; the sampling you can see the you can say this is the bootstrap sampling distribution. Bootstrap sampling distribution for alpha and beta ok. And we can see that 0 is somewhere here.

So, this is pretty much including the 0 the distribution and we can see that 1 is somewhere here. So, beta definitely concluding including 1 whereas, and alpha distribution including 0. So, it is fairly priced and the confidence interval for beta says that most likely it is as much risky as overall market.

(Refer Slide Time: 28:08)

15 of 20 Residual Bootstrap Regression

OLS Estimates of alpha and beta

	Estimate	Std. Error	t value	Pr(> t )
alpha	-0.0009	0.0010	-0.8790	0.3809
beta	1.0150	0.1157	8.7748	0.0000

-----




Residual Bootstrap Estimates of alpha and beta

	Estimate	Std. Error	2.5%	97.5%
alpha	-0.0009	0.001	-0.0030	0.0009
beta	1.0225	0.111	0.7968	1.2289

-----

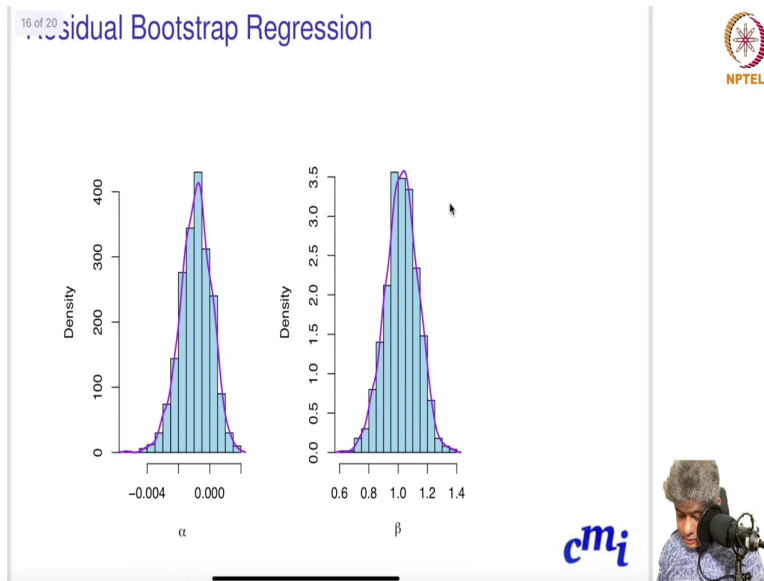
Paired Bootstrap Estimates of alpha and beta

	Estimate	Std. Error	2.5%	97.5%
alpha	-0.0009	0.0010	-0.0030	0.0009
beta	1.0137	0.0915	0.8415	1.2004



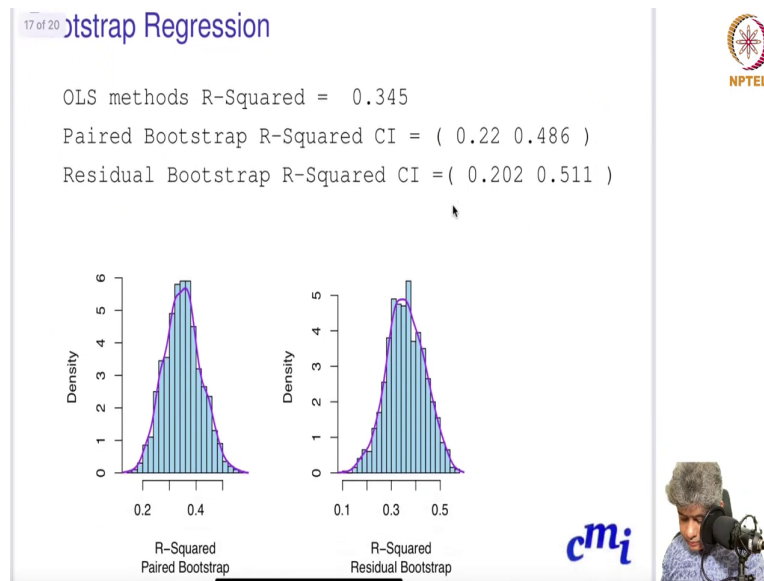
Here is we found the residual bootstrap regression also. What we found that residual bootstrap alpha is pretty much same, even the confidence interval also pretty much similar and paired bootstrap because one of the reason is both we found that the in CAPM when we studied this that homoscedasticity was ok assumption. So, naturally residual bootstrap will be fine. So, naturally paired bootstrap and residual bootstrap tend to agree on each other ok.

(Refer Slide Time: 28:47)





(Refer Slide Time: 28:52)



So, here is the residual bootstrap regression the similar kind of distribution we are finding here. Here is the here is another additional advantage of bootstrap regression technique. So, in OLS method if you just use OLS method with the normality assumption on the residual the R squared turns out to be 0.345 ok.


Now, because you have bootstrap samples in each bootstrap sample you can calculate the R squared and so you can find a confidence interval for R squared which regular of the shelf you know regression analysis cannot produce. So, of the shelf regression analysis cannot produce the R regular R squared. The confidence interval for regular R squared, but bootstrap regression can give you a confidence interval for R squared ok. So, this is an advantage of bootstrap regression idea.


(Refer Slide Time: 29:59)

The idea of Bootstrap Statistics

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

- ▶ Random Forest (2002-2003)
- ▶ Ensemble model
- ▶ Bagging etc.

cm: 



So, the idea of bootstrap statistics or resampling technique can be found in idea of random forest, ensemble model, bagging. So, in the machine learning technique bootstrap statistics is almost everywhere in many many techniques are being kind of you know inspired by the bootstrap statistics. Random forest was I think around 2002 or 2003 was developed ensemble model was also in that time bagging followed by bagging.

(Refer Slide Time: 30:45)

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

- ▶ Random Forest (2002-2003)
- ▶ Ensemble model
- ▶ Bagging etc.

Thank you.  
Harendra on

cm

NPTEL

The slide features a list of machine learning techniques associated with bootstrap statistics. The text is handwritten in pink. A speaker is visible in the bottom right corner of the video frame, and the NPTEL logo is in the top right.

So, with that I will stop here, thank you very much see you in the next video with hands on.

Thank you, bye.