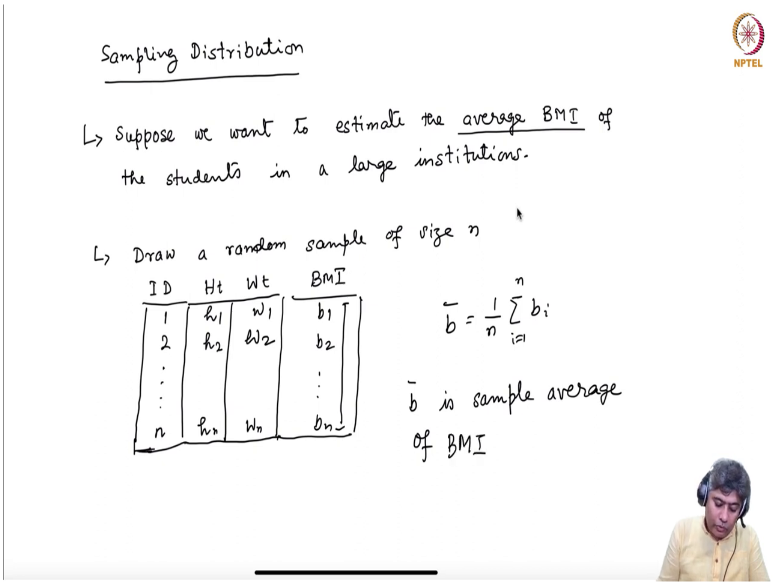
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 15
Sampling Distribution and Statistical Inference of Regression Coefficient

Welcome back to the part B of lecture 4 and now we are going to start Sampling Distribution of Regression Coefficient. However, before we start sampling distribution of regression coefficient we should little bit discuss about what is sampling distribution. If you already know what is sampling distribution my recommendation is you can skip this part and but if you were not so, sure if you are aware of what is sampling distribution then please continue watching this video.

(Refer Slide Time: 01:13)



Sampling Distribution

↳ Suppose we want to estimate the average BMI of the students in a large institutions.

↳ Draw a random sample of size n

ID	Ht	Wt	BMI
1	h_1	w_1	b_1
2	h_2	w_2	b_2
⋮			⋮
n	h_n	w_n	b_n

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i$$

$\bar{b}$ is sample average of BMI

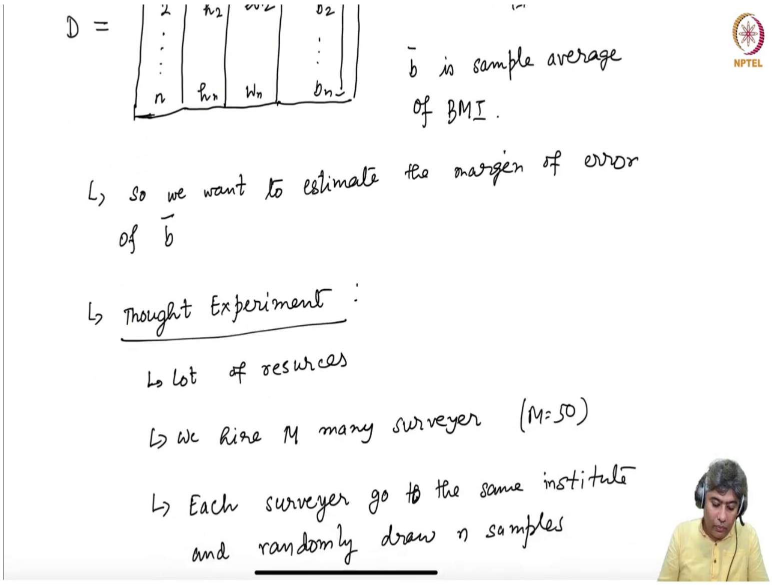
So, what we will do? We will try to understand what is the sampling distribution. So, sampling distribution is an very interesting concept suppose you suppose we want to; we want to estimate the average body mass index of the students of the students in a large institution intuitions ok.

Now, what we will do? We will typically what we will do we will draw a random sample of size n. So, from that large institution draw random sample of size n ok. So, you have probably this kind of data set like maybe height and weight and ID maybe this is first sample whatever the height whatever the weight this is the second sample second drawn from the you know from the population is and the height and weight of the second sample.

And then nth sample height and weight and using height and weight we can calculate the body mass index say suppose this is b 1 b 2 and b n. What we are interested in to calculate the average you know body mass index. So, we are we are interested in average body mass index of the students.

So, we can simply take b bar the sample mean of this values ok this body mass indices 1 by n summation b i i equal to 1 to n. Now b bar is sample average of body mass intakes. So, we want to estimate.

(Refer Slide Time: 04:44)



$$D = \begin{array}{|c|c|c|c|} 2 & h_2 & w_2 & b_2 \\ \vdots & & & \vdots \\ n & h_n & w_n & b_n \end{array}$$

$\bar{b}$ is sample average of BMI.

↳ So we want to estimate the margin of error of $\bar{b}$

↳ Thought Experiment :

↳ Lot of resurces

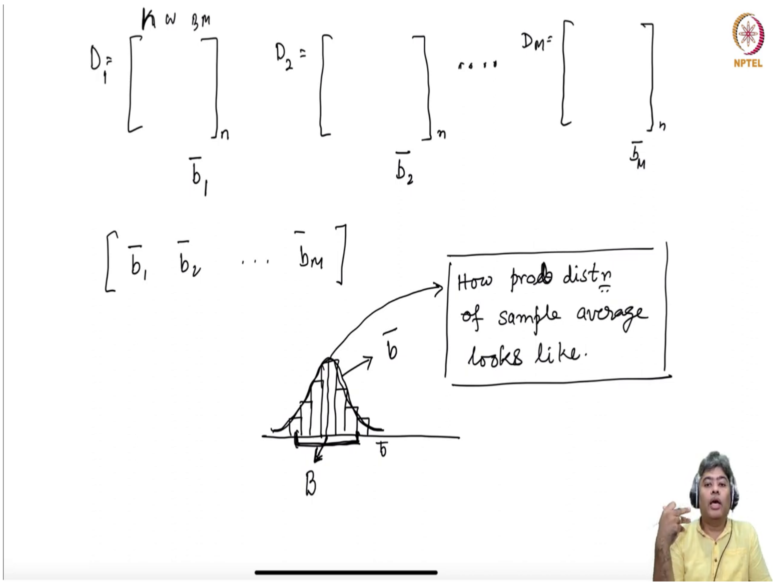
↳ We hire M many surveyer (M=50)

↳ Each surveyer go to the same institute and randomly draw n samples

So, we want to estimate the margin of error; margin of error for of b bar sample mean. Now, what we will do? We will sort of a do a thought experiment ok we will do a thought experiment we will do a thought experiment ok. So, what is thought experiment? The thought experiment is suppose you have infinite resources suppose you have infinite resources lot of resources ok not infinite I would say lot of resources.

You hire you we hire maybe 100 M many surveyors each surveyors M maybe 50 ok each surveyor go to the same institute and randomly draw n samples ok. Now randomly draw n samples what happens is if previously in real life we will have only one sample one data set that is it and we have to do all the analysis based on this one data set.

(Refer Slide Time: 07:26)



How prob distn of sample average looks like.

Now, what we will do? What we will do? We have M many data set we have M many data set the first D 1 D 2 D M and each data set has computed height weight and body mass index each data set they have collected height and weight and from there they calculated the body mass index.

Now, what you can do you instead of one data set you have M many data set or M if M is 50 they have 50 different data set. So, now, you can calculate the sample mean from the data set 1, sample mean of the body mass index from data set 2, sample mean from the body mass index from the Mth data set.

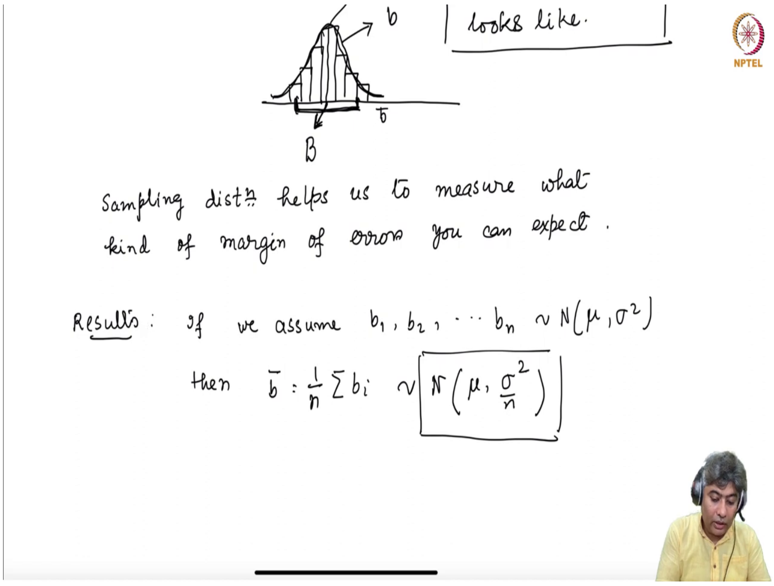
Remember that each of them have n samples ok. So, now, you have this b 1 bar, b 2 bar, b m bar now this vector you can take and draw a histogram of body mass index. So, this histogram this histogram is the histogram of essentially b bars. So, this histogram tells us how the distribution of sample average looks like.

So, b bar this distribution tells us how probability distribution probability distribution of sample average looks like ok. So, this is a interesting concept why? Because you can say that ok now since the sampling distribution of b bars looks like this.So, we can say that maybe original average b bar B is somewhere the body mass index expected body mass index is somewhere here.

And with the some you know confidence interval like you can give a sort of a range that most likely this is where the true value will looks be there and if you get a b bar here then it is likely that it will be far away from the. So, you have a sense that whether a sample mean is too far away or reasonably close. So, that kind of probabilistic statement now you can make.

So, sampling distribution enable us to make a probabilistic statement about our estimation it gives us to calculate what kind of margin of error we can expect.

(Refer Slide Time: 11:20)



Sampling distn helps us to measure what kind of margin of error you can expect.

Results: If we assume $b_1, b_2, \cdots b_n \sim N(\mu, \sigma^2)$

then $\bar{b} = \frac{1}{n}\sum b_i \sim \boxed{N\left(\mu, \frac{\sigma^2}{n}\right)}$

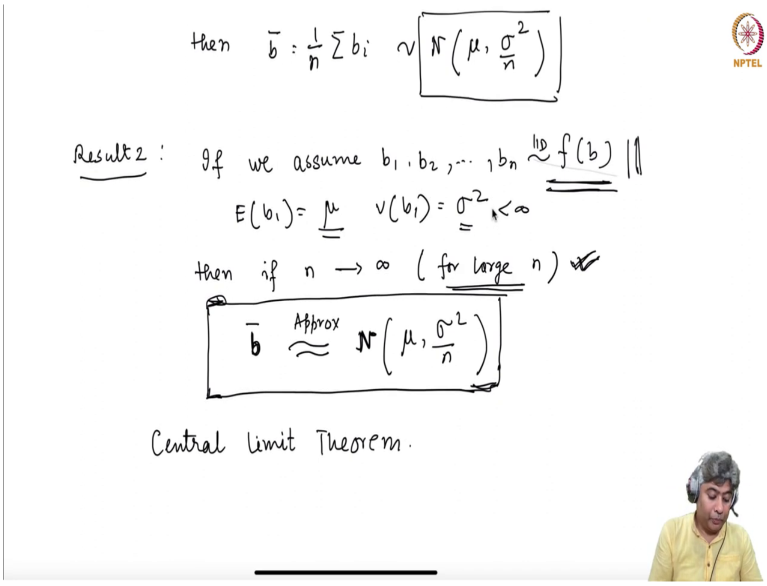
So, sampling distribution sampling distribution helps us to identify to measure to measure what kind of margin of error margin of error you can expect. Now given this thing if this is the thing now this was only for sample mean, but in reality we cannot have m many surveyors this is too much this is remember that this is a thought experiment.

This was not a you know real life in a situation, we cannot have M many samples we all we have in reality we have this one sample this one sample that set we do not have more than one sample. So, we have to calculate this margin of error using the whatever one sample that we have.

So, turns out that few results few results if you assume if we assume that b 1, b 2 dot dot dot b n they follow normal distribution with some mean mu and some variance sigma square then b bar which is sample mean 1 by n summation b i will follow normal distribution with mean mu and variance sigma square by n.

So, this is the sampling distribution of b bar will follow normal distribution with mean mu and variance sigma square by n. So, this is the first result and this is result 1.

(Refer Slide Time: 13:58)



then $\bar{b} = \frac{1}{n}\sum b_i \sim \boxed{N\left(\mu, \frac{\sigma^2}{n}\right)}$

Result 2 : If we assume $b_1, b_2, \cdots, b_n \overset{\text{iid}}{\sim} f(b)$

$E(b_1) = \mu \quad V(b_1) = \sigma^2 < \infty$

then if $n \to \infty$ (for large n)

$$\boxed{\bar{b} \overset{\text{Approx}}{\approx} N\left(\mu, \frac{\sigma^2}{n}\right)}$$

Central Limit Theorem.

And then there is another very interesting result 2 that if we assume; if we assume b 1, b 2, dot dot dot b n they are independent and identically distributed some distribution with some pdf which with some mean expectation of b 1 is mu and variance of b 1 is sigma square ok.

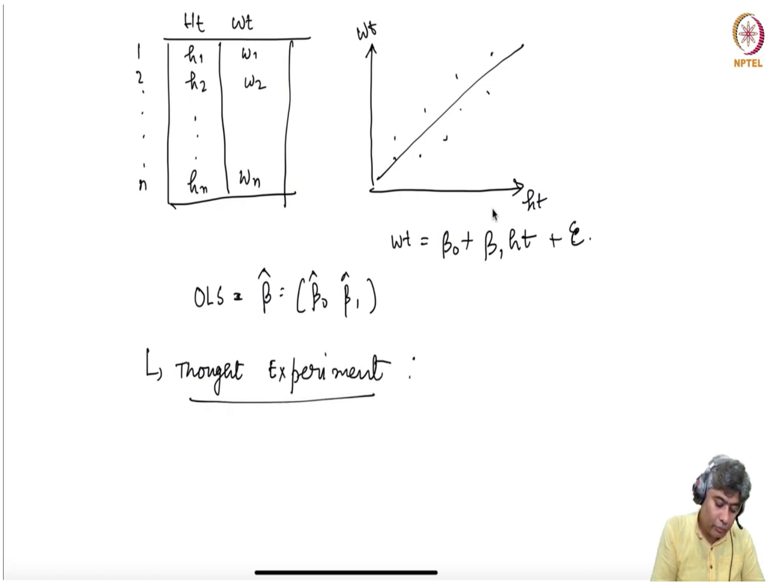
Then if n goes to infinity; that means, if you are for large n for large n the large sample size, the b bar will approximately follow normal distribution with mu and sigma square by n. So, what does it mean? It means, basically your BMI can follow any other distribution it can follow say lognormal distribution or some gamma distribution does not matter, I am saying it follow any distribution with the proper probability density function pdf or probability function it may follow binomial distribution all you know.

Some proper probability distribution it follows with finite mean and finite variance ok. Finite mean and finite variance then if you ensure you have large enough sample then the sample mean the sampling distribution of sample mean will approximately follow normal distribution and this is a huge result.

This is a huge result and this result is known as central limit theorem this result is known as central limit theorem ok. This is a huge result in a sense if you know if you even if you do not have to assume any distribution on the f b or like you know if b could be any distribution, it could be binomial or (Refer Time: 16:42) or you know you can it can be you know gamma lognormal anything any distribution.

But still the all you have to ensure that the sample size is n and then all you have to do is b is following the sample mean b bar still follow will approximately follow normal distribution and this result is known central limit theorem. Now, how can we conceptualize the sampling distribution of regression coefficient?

(Refer Slide Time: 17:25)



	Ht	wt
1	h_1	w_1
2	h_2	w_2
.	.	
.	.	
n	h_n	w_n

$wt = \beta_0 + \beta_1 ht + \varepsilon.$

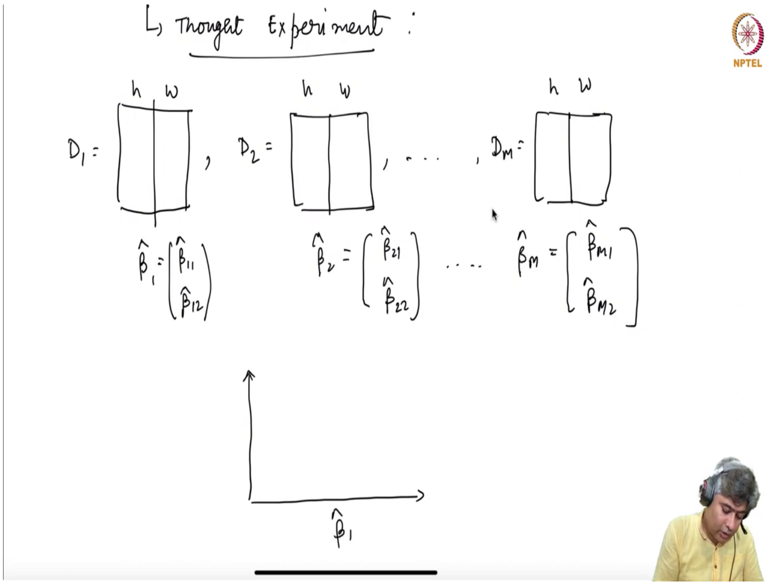
$OLS = \hat{\beta} : (\hat{\beta}_0 \ \hat{\beta}_1)$

↳ Thought Experiment :

So, we have the data like this height weight then suppose this is the data that we have 1, 2 up to n we have h 1 w 1, h 2 w 2, dot dot dot h n w n and we want to say from the height you want to say something about the weight. Suppose, this is the data set that we have. So, we want to fit a straight line like this. So, we want to fit a wt as a function of say beta naught plus beta 1 height plus some error ok.

Now what we can think of is we can use this data set using OLS estimator we can compute the beta hat as beta naught hat beta 1 hat ok. And then what we can do? We can think of again do the thought experiment we can do a thought experiment ok.

(Refer Slide Time: 18:56)



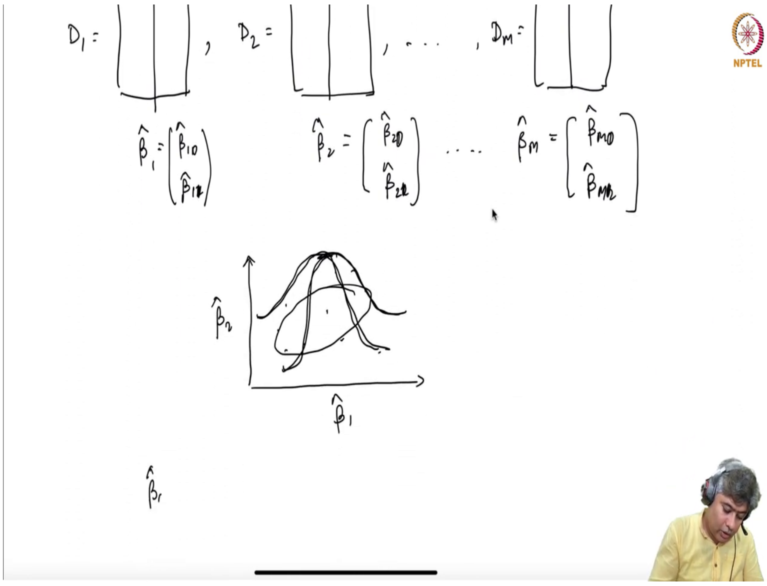
↳ Thought Experiment :

$\hat{\beta}_1 = \begin{pmatrix} \hat{\beta}_{11} \\ \hat{\beta}_{12} \end{pmatrix}$ $\hat{\beta}_2 = \begin{bmatrix} \hat{\beta}_{21} \\ \hat{\beta}_{22} \end{bmatrix}$ . . . . $\hat{\beta}_M = \begin{bmatrix} \hat{\beta}_{M1} \\ \hat{\beta}_{M2} \end{bmatrix}$

In the thought experiment what we can do? We can have m many samples. So, D 1 it is like height and weight, then D 2 another data set another surveyor go to the field and collect the samples height and weight and Mth surveyor go to the field and collect the. Now, what we can do from each thing we can compute the beta hat beta 2 hat. So, beta hat is beta 11 hat beta 12 hat.

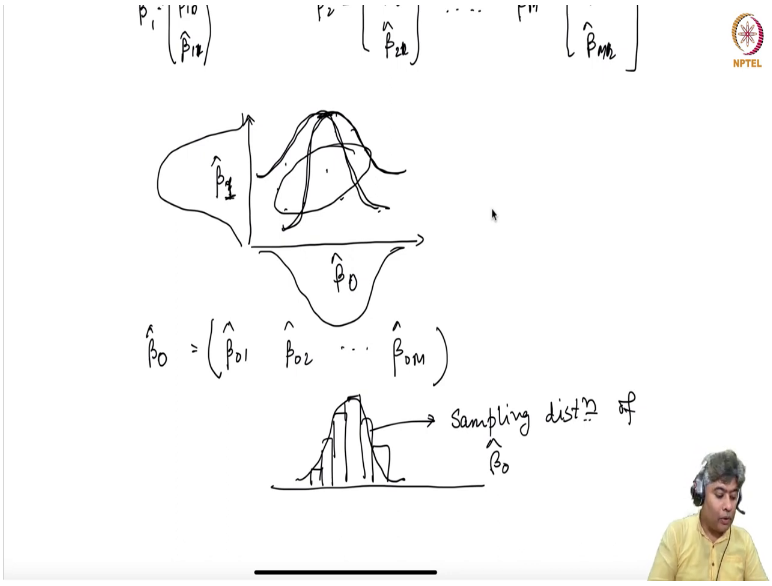
Now beta 2 hat is beta 21 hat, beta 22 hat from the Mth data set we can compute the Mth M. So, beta M 1 hat, beta M 2 hat. Now, what we can do? We can just plot these think of this guys as here it is beta 1 hat, here it is beta 2 hat and we can if we just plot this guys.

(Refer Slide Time: 20:36)


$D_1 =$, $D_2 =$, . . . , $D_M =$
$\hat{\beta}_1 = \begin{pmatrix} \hat{\beta}_{10} \\ \hat{\beta}_{11} \end{pmatrix}$
$\hat{\beta}_2 = \begin{bmatrix} \hat{\beta}_{20} \\ \hat{\beta}_{21} \end{bmatrix}$ . . . .
$\hat{\beta}_M = \begin{bmatrix} \hat{\beta}_{M0} \\ \hat{\beta}_{M1} \end{bmatrix}$
$\hat{\beta}_2$
$\hat{\beta}_1$
$\hat{\beta}_1$

So, we have essentially a two dimensional distribution and this two dimensional distribution is essentially the joint distribution of sampling distribution of beta hat ok. This is the sampling distribution of beta hat if we just take the beta 1 hat, the intercept for the intercept. So, ok we have used a 0 here. So, we can use 0 1 0 1 and 0 1.

(Refer Slide Time: 21:26)

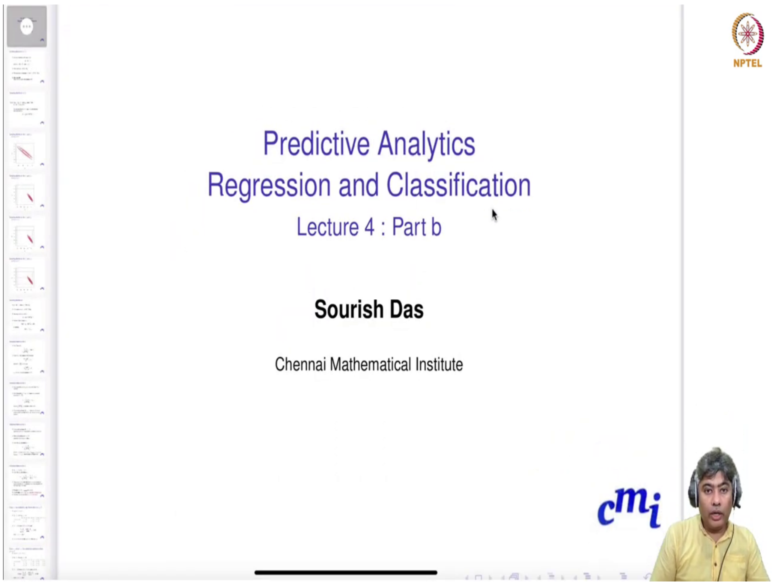


$\hat{\beta}_0 = (\hat{\beta}_{01} \quad \hat{\beta}_{02} \quad \cdots \quad \hat{\beta}_{0M})$

So, here it will be 0 and here it will be 1. So, now, if we just intercept beta 0 hat. So, all we have beta 0 1 hat, beta 0 2 hat dot dot dot beta 0 M hat. So, I from m data set I have M intercepts we can just draw the histogram and this distribution will be the sampling distribution of intercept. So, this is how the distribution will look like the sampling distribution of beta intercept beta 0 hat and this will be the sampling distribution of the slope.

So, that is how we are going to discuss the this is the sampling distribution of regression coefficient.

(Refer Slide Time: 22:32)



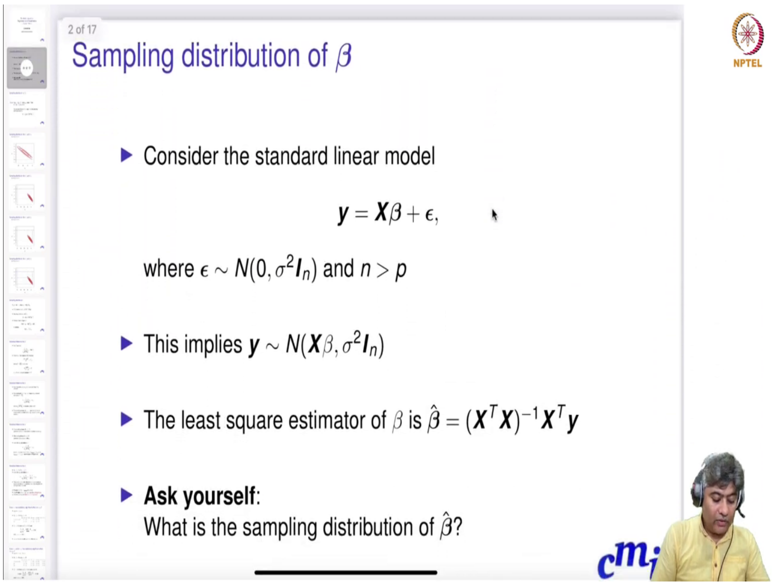
Predictive Analytics
Regression and Classification
Lecture 4 : Part b

Sourish Das

Chennai Mathematical Institute

Now we will go back and we will start with the regression analytics of part b.

(Refer Slide Time: 22:41)



2 of 17

Sampling distribution of β

- Consider the standard linear model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon,$$

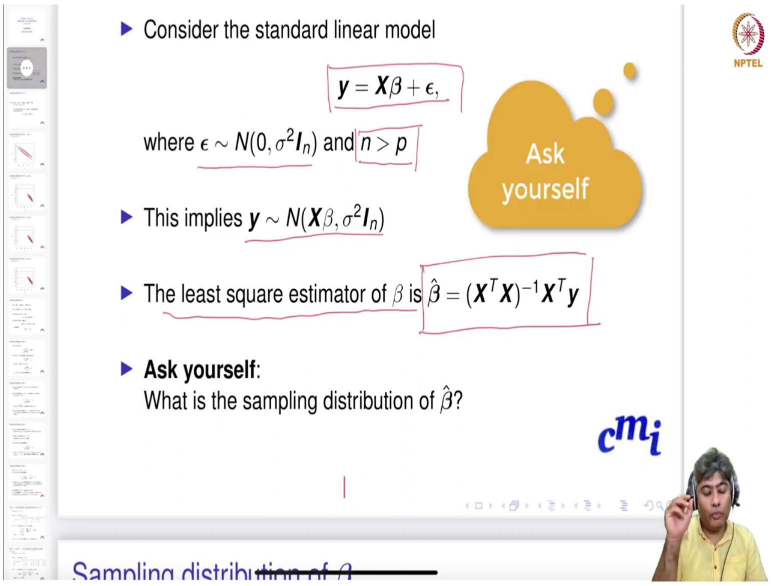
where $\epsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$ and $n > p$

- This implies $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$
- The least square estimator of β is $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- **Ask yourself**:
  What is the sampling distribution of $\hat{\beta}$?

So, now, consider our standard linear model ok in the standard linear model what we have is y equal to X beta plus epsilon the standard model that we have. And then where epsilon follow normal 0 sigma square I n and n is greater than p. This is we always making sure that n is greater than p because if n is not greater than p then the it is not a OLS estimator is not going to be full rank. So, we will not have a proper estimate. So, this is required.

Now, this immediately implies y equal to X beta normal X beta sigma square I n. From here we can we also know that least square estimator of beta hat is X transpose X inverse X transpose y. So, the question is that ask yourself that what is the sampling distribution of beta hat?

(Refer Slide Time: 24:01)



- Consider the standard linear model

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon,}$$

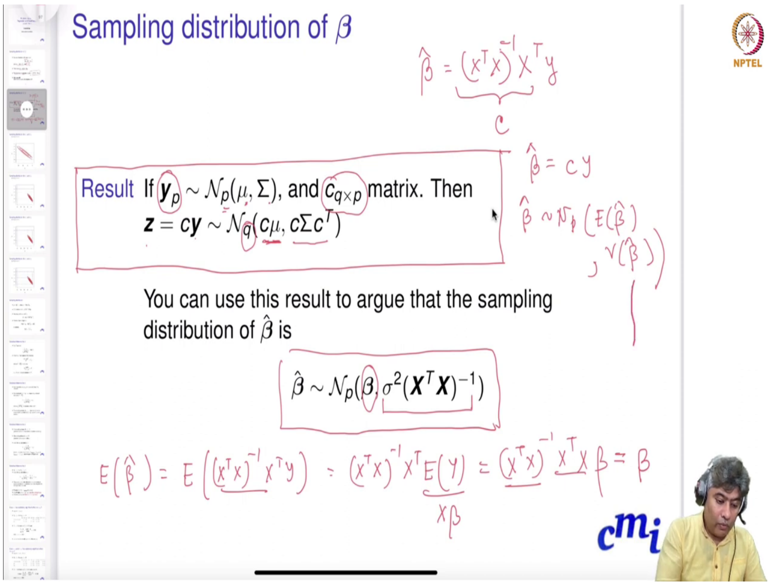
where $\epsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$ and $\boxed{n > p}$

- This implies $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$
- The least square estimator of β is $\boxed{\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$
- **Ask yourself**:
What is the sampling distribution of $\hat{\beta}$?

Sampling distribution of β

So, my recommendation is you stop yourself, you pause your video for about 5 minutes try to figure out what is the sampling distribution of regression coefficient or the sampling distribution of beta hat in this case and see if the results match with my what I am going to tell next.

(Refer Slide Time: 24:25)



Sampling distribution of β

$$\hat{\beta} = \underbrace{(X^TX)^{-1}X^T}_{c} y$$

$$\hat{\beta} = cy$$

$$\hat{\beta} \sim N_p\left(E(\hat{\beta}), V(\hat{\beta})\right)$$

$$\boxed{\text{Result If } y_p \sim \mathcal{N}_p(\mu, \Sigma), \text{ and } c_{q\times p} \text{ matrix. Then } z = cy \sim \mathcal{N}_q(c\mu, c\Sigma c^T)}$$

You can use this result to argue that the sampling distribution of $\hat{\beta}$ is

$$\boxed{\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X^TX)^{-1})}$$

$$E(\hat{\beta}) = E\left((X^TX)^{-1}X^Ty\right) = (X^TX)^{-1}X^T\underbrace{E(y)}_{X\beta} = (X^TX)^{-1}X^TX\beta = \beta$$

I hope you have find the results by now. So, what I am going to do? I am going to use this result of multivariate Gaussian distribution or multivariate normal distribution. If you have a vector of size p which follow p variant normal with mean vector mu and covariance matrix sigma.

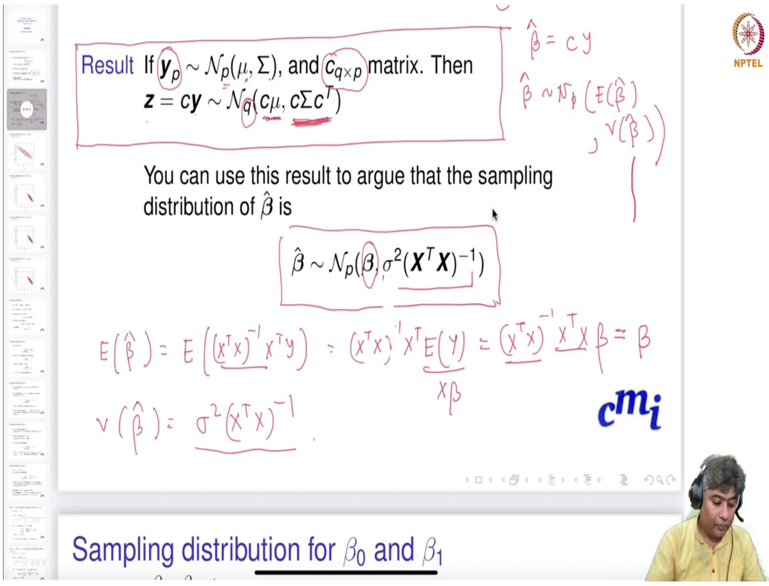
Now, you have a some coefficient matrix c which is of q cross p dimension then c times y will give you a new variable z and this will still follow will follow q dimension normal with mean c mu and variance c sigma c transpose. You can use this result to argue that the sampling distribution of beta hat follow p variant normal with mean beta itself the true beta and the covariance sigma squared X transpose X inverse how can you do that?

Now, you see beta hat is X transpose X inverse X transpose y correct? This X transpose X you can treat it as a c. So, you can write beta hat as c times y then expectation of beta hat will be. So, naturally beta hat will follow if you write beta hat to be c times y then you can write beta hat follow some normal distribution here in this case it will be p variant normal with some expectation of beta hat and variance of beta hat.

So, this will follow normal for sure you have to only figure out what is the expected value of beta hat. Expected value of beta hat will be expected value of X transpose X inverse X transpose y which is c. So, the c comes out then it will be X transpose X inverse X transpose expected value of y.

Now, what is expected value of y? This is X beta this is X beta. So, I have X transpose X inverse X transpose X beta. Now this is X transpose X this is X transpose X inverse. So, this will yield a identity matrix. So, we left with beta.

(Refer Slide Time: 27:43)



Result If $\boldsymbol{y}_p \sim \mathcal{N}_p(\mu, \Sigma)$, and $c_{q\times p}$ matrix. Then $\boldsymbol{z} = c\boldsymbol{y} \sim \mathcal{N}_q(c\mu, c\Sigma c^T)$

You can use this result to argue that the sampling distribution of $\hat\beta$ is

$$\boxed{\hat\beta \sim \mathcal{N}_p(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})}$$

$\hat\beta = cy$

$\hat\beta \sim N_p(E(\hat\beta), V(\hat\beta))$

$E(\hat\beta) = E((X^TX)^{-1}X^Ty) = (X^TX)^{-1}X^T\underbrace{E(y)}_{X\beta} = (X^TX)^{-1}X^TX\beta = \beta$

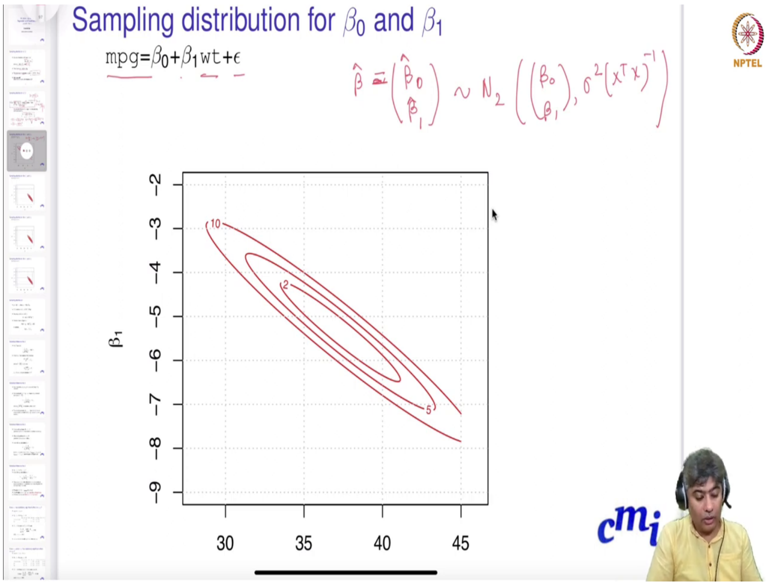
$V(\hat\beta) = \sigma^2(X^TX)^{-1}$

Sampling distribution for β_0 and β_1

Now you have to figure out you should figure out that how variance of beta hat will be sigma squared X transpose X inverse. I am leaving it for you to figure out it is very simple all you have to do just apply variant this you have to just apply this result that variance of c y is c sigma c transpose.

If you apply this that will be good enough that will yield you this result sigma square X transpose X inverse all right.

(Refer Slide Time: 28:19)



Sampling distribution for β_0 and β_1

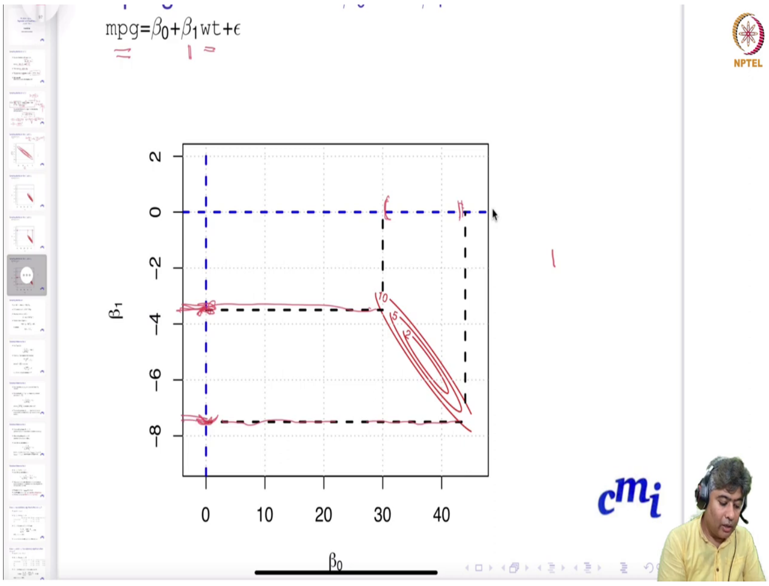
mpg=β_0+$\beta_1$wt+ϵ

$\hat{\beta} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} \sim N_2\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \sigma^2(X^TX)^{-1}\right)$

Next how we are now taking the sampling distribution of beta naught and beta 1? So, we are considering this model m p g as a function of beta naught plus beta 1 times weight plus epsilon and we have figured out that beta hat follow in this case beta hat will be essentially beta 1 hat and beta beta naught hat and beta 1 hat.

So, this will follow a bivariate normal with beta naught and beta 1 as mean with some sigma square X transpose X inverse that you have to we have figured out last part of the lecture. So, that is how. So, since it is a bivariate normal this is going to be the. So, on the x axis I have put beta naught on the y axis I have put the beta 1 and the sampling distribution behaves like this ok.

(Refer Slide Time: 29:47)



Now, what I am going to do? I am just taking this same distribution and just including 0 in this picture. So, previously you see my x axis was somewhere between 28 to 45 and minus 9 to minus 2 I am just expanding the x axis and y axis to include 0. So, now, I am it is from minus 10 to 2 and it is around including may be minus 1 to 45.

So, you can see that this distribution is somewhere here. Now what we can think of this 95 percent of the probability mass for beta naught is somewhere in this region which is far from 0 here is 0 and 95 percent of the mass of beta naught actually going to be in this region.

Similarly, 95 percent of the mass of beta 1 is going to be in this region ok in minus 8 and minus 6. So, that is how we now we can say that beta naught is far from 0 similarly beta 1 is far from 0 it has a negative values very likely 95 percent probability may be 99 percent probability that beta 1 is negative and it has a strong negative correlation.

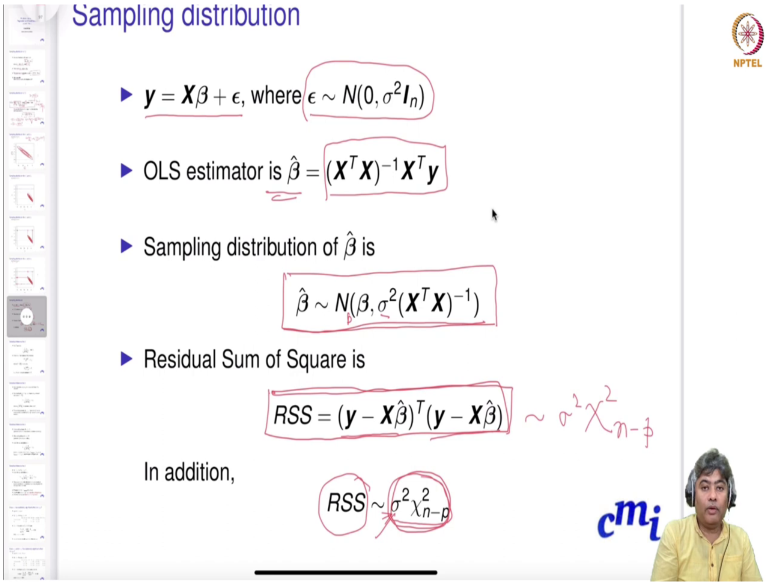
Now, what is beta 1? Beta 1 is the slope of weight and miles per gallon ok so; that means, if I am very confident that weight has a negative correlation with miles per gallon and because beta 1 is most likely with 95 percent probability it is going to take negative value.

(Refer Slide Time: 31:56)

NPTEL
Sampling distribution is foundation of statistical inference

This kind of probabilistic inference you can do with sampling distribution of regression coefficient. Therefore, sampling distribution of regression coefficient is absolutely foundationally fundamental for doing any statistical inference.

(Refer Slide Time: 32:22)



Sampling distribution

- $y = X\beta + \epsilon$, where $\boxed{\epsilon \sim N(0, \sigma^2 I_n)}$
- OLS estimator is $\hat{\beta} = \boxed{(X^T X)^{-1} X^T y}$
- Sampling distribution of $\hat{\beta}$ is

$$\boxed{\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})}$$

- Residual Sum of Square is

$$\boxed{RSS = (y - X\hat{\beta})^T (y - X\hat{\beta})} \sim \sigma^2 \chi^2_{n-p}$$

In addition, $RSS \sim \sigma^2 \chi^2_{n-p}$

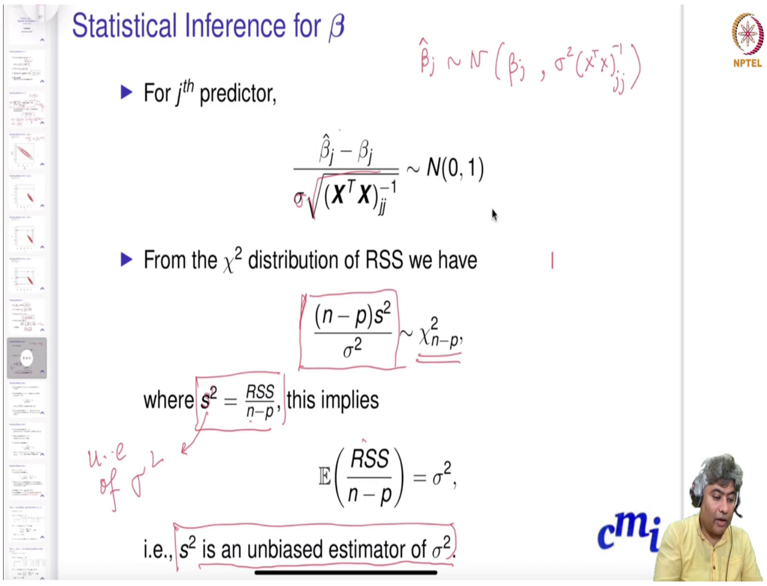
So, now we have figured out that y equal to X beta plus epsilon and epsilon for normal 0 sigma square I n our OLS estimator is beta hat which X transpose X inverse X transpose y and beta hat follow normal the sampling distribution of beta hat is normal p variant normal with mean as beta itself that a coefficient itself with the sigma square and x transpose sigma square times X transpose X inverse.

Now, residual sum of square RSS y minus X beta hat transpose times y minus X beta hat is the residual sum of square and it is found that residual sum of square follow scaled chi squared distribution scaled by sigma square and this is helpful because; that means, if I have to do any statistical inference on sigma square; on sigma square.

Then we can use residual sum of square to do that residual if we consider residual sum of square as a statistic a particular statistic see it involves only data x y x and beta hat and from we know the distribution, the distribution is sigma square chi square with n minus p degrees of freedom.

So, using this information we can do a statistical inference for sampling variance or residual variance.

(Refer Slide Time: 34:05)



Statistical Inference for β

$\hat{\beta}_j \sim N\left(\beta_j , \sigma^2 (X^T X)^{-1}_{jj}\right)$

- For j^{th} predictor,

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} \sim N(0,1)$$

- From the χ^2 distribution of RSS we have

$$\boxed{\frac{(n-p)s^2}{\sigma^2}} \sim \underline{\underline{\chi^2_{n-p}}},$$

where $\boxed{s^2 = \frac{RSS}{n-p}}$, this implies

u.e of σ^2

$$\mathbb{E}\left(\frac{RSS}{n-p}\right) = \sigma^2,$$

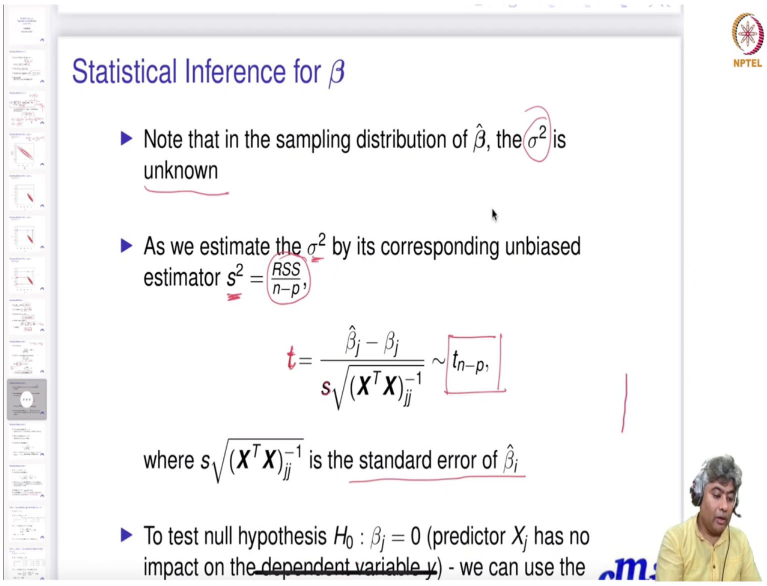
i.e., $\boxed{s^2 \text{ is an unbiased estimator of } \sigma^2}$.

Now an interesting thing that is how to do statistical inference for beta? So, for jth predictor in this case what we will have? We have we know that beta j hat since beta j hat will follow normal beta j and it will be like sigma square the ith element of X transpose X inverse the or the jth element of the jth element of x transpose x inverse.

If we do that then essentially we can write beta j hat minus beta j sigma times square root of X transpose X inverse j jth element will follow normal 0 1 and from chi squared distribution we know that n minus p a square by sigma square will follow chi squared with n minus p degrees of freedom. So, s square which is residual sum of square by n minus p what is it implies? It implies expected value of residual sum of square by n minus p is equal to sigma square.

So, this means s square is an unbiased estimator of sigma square. So, s square is an unbiased estimator of sigma square. So, this is an interesting result that we have.

(Refer Slide Time: 35:57)



Statistical Inference for β

- Note that in the sampling distribution of $\hat{\beta}$, the σ^2 is unknown
- As we estimate the σ^2 by its corresponding unbiased estimator $s^2 = \frac{RSS}{n-p}$,

$$t = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} \sim \boxed{t_{n-p}},$$

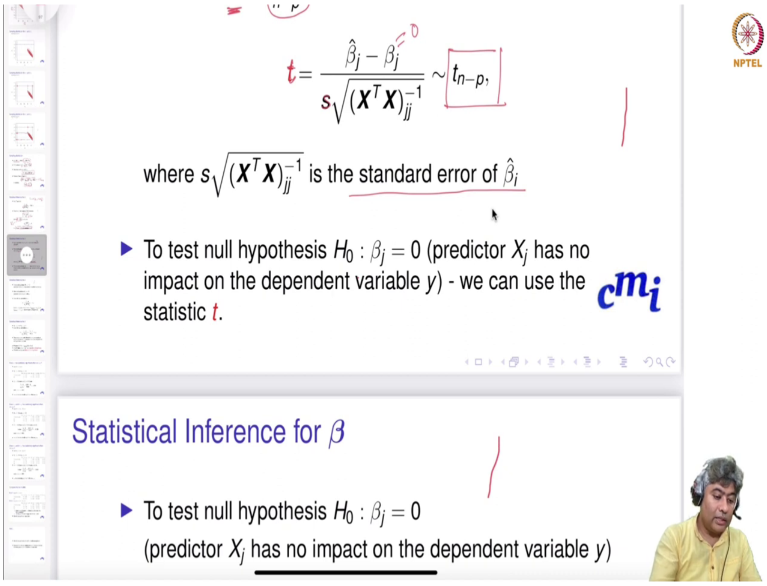
where $s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}$ is the standard error of $\hat{\beta}_i$

- To test null hypothesis $H_0 : \beta_j = 0$ (predictor X_j has no impact on the dependent variable y) - we can use the

So, note that in the sampling distribution of beta hat the sigma square is unknown. Since sigma square is unknown. So, we have to estimate 1one way to do that we estimate sigma square by the corresponding unbiased estimator of sigma square s square which is residual sum of square by n minus p.

So, if you estimate the with that then t divided by beta j hat minus beta j now if you now estimate sigma by s, then this will this t statistic will follow t distribution with n minus p degrees of freedom. So, where s is the standard error of the beta j hat.

(Refer Slide Time: 36:56)



$$t = \frac{\hat{\beta}_j - \overset{0}{\beta_j}}{s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} \sim \boxed{t_{n-p}},$$

where $s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}$ is the standard error of $\hat{\beta}_i$

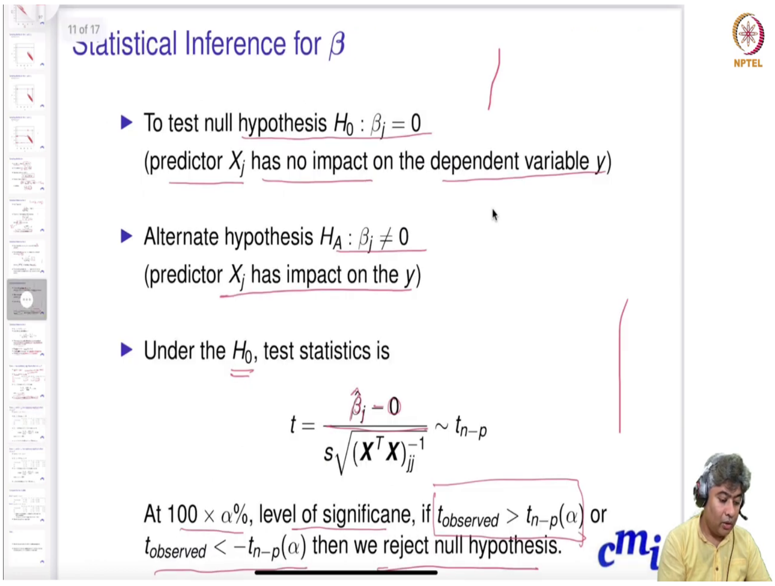
- To test null hypothesis $H_0 : \beta_j = 0$ (predictor X_j has no impact on the dependent variable y) - we can use the statistic t.

Statistical Inference for β

- To test null hypothesis $H_0 : \beta_j = 0$ (predictor X_j has no impact on the dependent variable y)

Now we can do the test the null hypothesis beta j equal to 0 what we have to do? We have to we are here we have to just put beta j equal to 0 and we will be all set. So, how we will do the statistical inference?

(Refer Slide Time: 37:11)



11 of 17

Statistical Inference for β

- To test null hypothesis $H_0 : \beta_j = 0$ (predictor X_j has no impact on the dependent variable y)
- Alternate hypothesis $H_A : \beta_j \neq 0$ (predictor X_j has impact on the y)
- Under the H_0, test statistics is

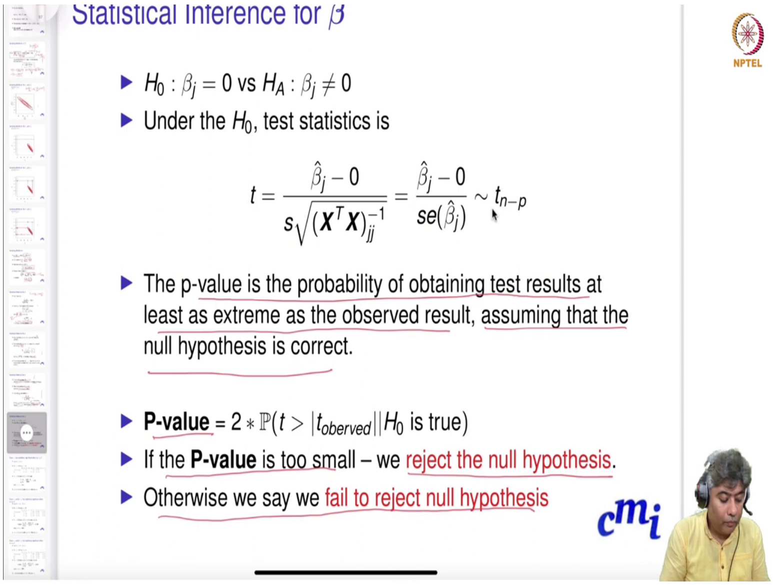
$$t = \frac{\hat{\beta}_j - 0}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-p}$$

At $100 \times \alpha\%$, level of significane, if $t_{observed} > t_{n-p}(\alpha)$ or $t_{observed} < -t_{n-p}(\alpha)$ then we reject null hypothesis.

We will do in this way. So, how can we do that? So, all we have to do test the null hypothesis X H naught beta j equal to 0 now what does it mean? If beta j equal to 0 that means, predictor X j has no impact on the dependent variable y that is what it means alternate hypothesis is beta j naught equal to 0 what is it means? It means, predictor X j has impact on y ok.

Now, under null hypothesis under the null hypothesis the test statistic will be beta j hat minus 0 divided by the standard error of the beta j hat.

(Refer Slide Time: 38:00)



Statistical Inference for β

- $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$
- Under the H_0, test statistics is

$$t = \frac{\hat{\beta}_j - 0}{s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}} = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} \sim t_{n-p}$$

- The p-value is the probability of obtaining test results at least as extreme as the observed result, assuming that the null hypothesis is correct.

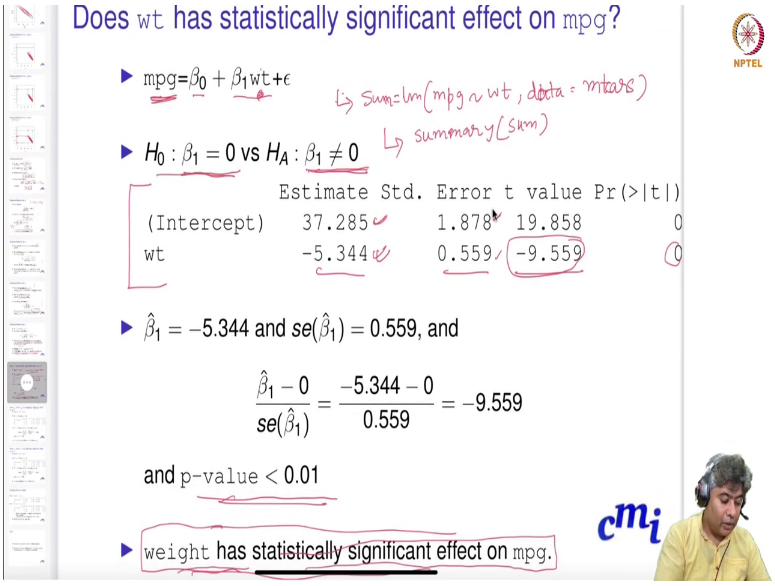
- **P-value** = $2 * \mathbb{P}(t > |t_{oberved}||H_0$ is true)
- If the **P-value** is too small – we reject the null hypothesis.
- Otherwise we say we fail to reject null hypothesis

So, what we will do and that follow t n minus 1 and at 100 into 1 minus alpha percent level of significance if the t observed is greater than t n minus p alpha or t observed is minus t n minus p alpha then we reject the null hypothesis. So, that is how we run the statistical inference for regression coefficient.

So, another way of doing it is simply instead of instead of looking at the t value, we can look into the P-value the P-value is the probability of obtaining the test result at least as extreme as observed result assuming that the null hypothesis is correct. So, we can calculate the P-value by multiplying probability of t greater than t observed under the null hypothesis by 2.

And then if the P-value is too small then we reject the null hypothesis otherwise we say we fail to reject the null hypothesis.

(Refer Slide Time: 39:13)



Does wt has statistically significant effect on mpg?

- mpg=$\beta_0 + \beta_1$wt+ϵ

↳ Sum=lm(mpg ~ wt, data = mtcars)

↳ summary(Sum)

- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     37.285      1.878  19.858        0
wt              -5.344      0.559  -9.559        0
```

- $\hat{\beta}_1 = -5.344$ and $se(\hat{\beta}_1) = 0.559$, and

$$\frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{-5.344 - 0}{0.559} = -9.559$$

and p-value < 0.01

- weight has statistically significant effect on mpg.

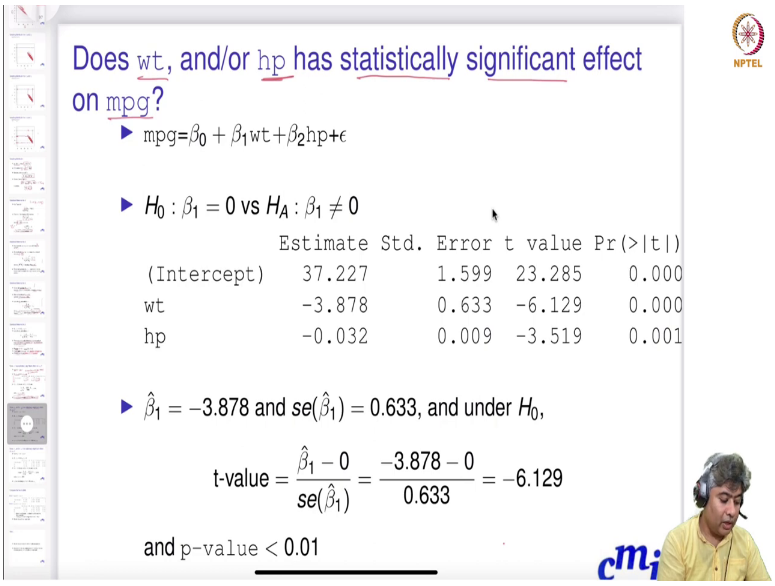
So, here is an example quick example are example that we have done. So, we run the my mt cars data set beta mpg beta naught plus beta 1 weight and so, null hypothesis is beta 1 equal to 0 versus beta 1 non zero. So, beta 1 equal to 0 means, null hypothesis is saying that weight does not have any effect on miles per gallon whereas, beta 1 non zero the alternative hypothesis is saying that beta 1 non zero; that means, we do have weights do have effect on the miles per gallon.

Now, the r if you run the r typically if you just run l m with m p g dollar weight with data equal to m t cars mt cars you have to put it in some object say call it summary and then print summary in sum if you run this two lines in r then you will give this you will get this output.

So, first is the estimate of the beta naught this is the estimate of the beta 1, this is the standard error we calculated and estimate divided by the standard error will give you the t value. So, for beta 1 you see the negative 9.559 is pretty small value and P-value is nearly 0 it is rounded off to 0.

So, we can say that since the P-value is really small weight has we can we will reject null hypothesis so; that means, weight has statistically significant effect on miles per gallon and that make sense. So, if you the if the car is large, it has a lot of weight naturally it fuel efficiency will be much less.

(Refer Slide Time: 41:29)



Does wt, and/or hp has statistically significant effect on mpg?

- mpg=$\beta_0+\beta_1$wt+$\beta_2$hp+ϵ
- $H_0:\beta_1=0$ vs $H_A:\beta_1\neq 0$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.227      1.599  23.285    0.000
wt            -3.878      0.633  -6.129    0.000
hp            -0.032      0.009  -3.519    0.001
```

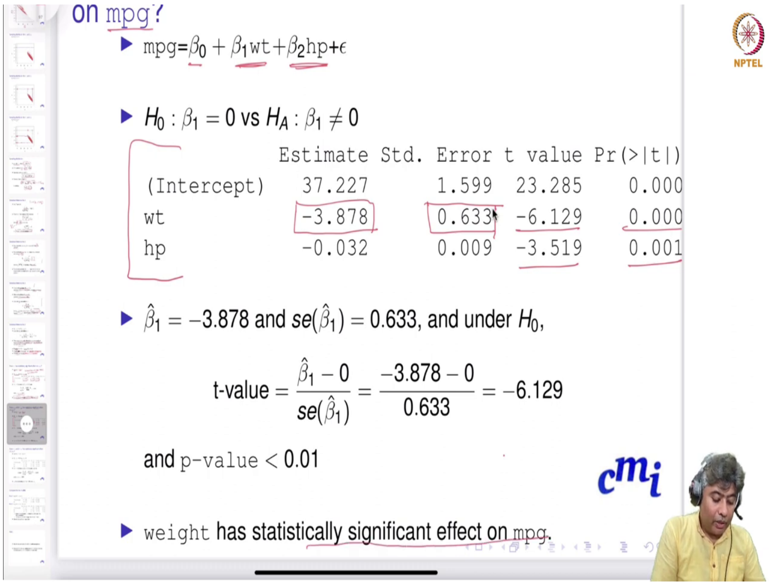
- $\hat{\beta}_1=-3.878$ and $se(\hat{\beta}_1)=0.633$, and under H_0,

$$\text{t-value}=\frac{\hat{\beta}_1-0}{se(\hat{\beta}_1)}=\frac{-3.878-0}{0.633}=-6.129$$

and p-value <0.01

Now, we look into another model. So, does weight and or horsepower two predictor model has statistically significant effect on miles per gallon. So, previously we were doing this we were trying this model sorry yeah this was my model we were trying this model.

(Refer Slide Time: 42:00)



on mpg?

- mpg=$\beta_0 + \beta_1$wt+$\beta_2$hp+ϵ
- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.227      1.599  23.285    0.000
wt            -3.878      0.633  -6.129    0.000
hp            -0.032      0.009  -3.519    0.001
```

- $\hat{\beta}_1 = -3.878$ and $se(\hat{\beta}_1) = 0.633$, and under H_0,

$$\text{t-value} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{-3.878 - 0}{0.633} = -6.129$$

and p-value < 0.01

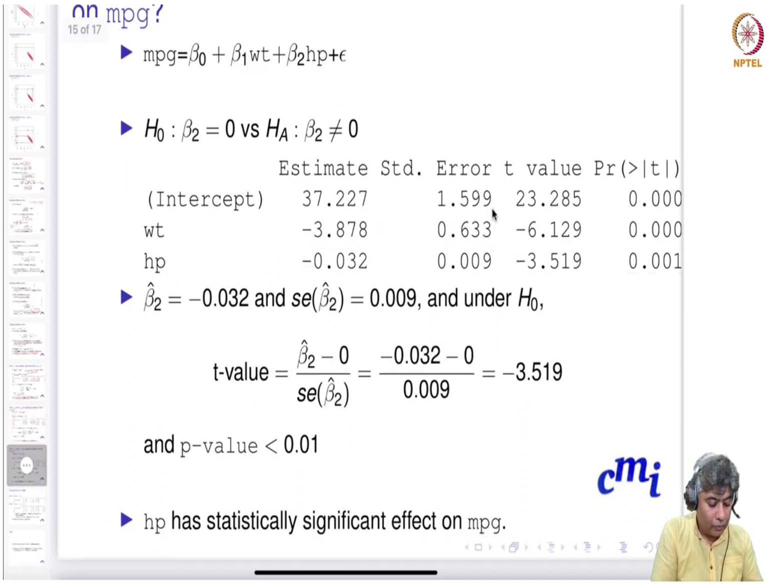
- weight has statistically significant effect on mpg.

Now we want beta naught plus beta 1 weight along with we want horsepower as the second independent variable and we run the same way we run the r and we get this estimates this result. So, now you can; obviously, because your model is different OLS will give you slightly different estimates.

So, previously the estimates of beta 1 was negative 5.344 now it is negative 3.878. So, the estimate has changed standard error if you look into the standard error standard error has was previously when you I have only one independent variable only weight that time it was 0559 now, it is 0.633.

So, the standard error has gone up ok, but still the t value is quite small for both weight and horsepower p values are also very small. So, weight has still have a significant effect on miles per gallon ok it has still have a significant effect, but what we are seeing that standard error of regression coefficient for weight or the standard error for beta 1 has slightly gone up.

(Refer Slide Time: 43:37)



on mpg?

15 of 17

- $\text{mpg}=\beta_0+\beta_1\text{wt}+\beta_2\text{hp}+\epsilon$
- $H_0 : \beta_2 = 0$ vs $H_A : \beta_2 \neq 0$

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.227      1.599  23.285    0.000
wt             -3.878      0.633  -6.129    0.000
hp             -0.032      0.009  -3.519    0.001
```

- $\hat{\beta}_2 = -0.032$ and $se(\hat{\beta}_2) = 0.009$, and under H_0,

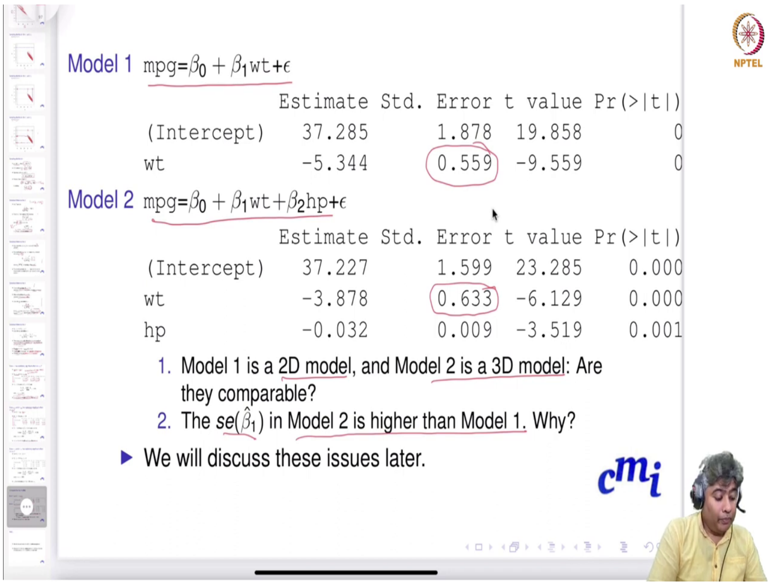
$$\text{t-value} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{-0.032 - 0}{0.009} = -3.519$$

and p-value < 0.01

- hp has statistically significant effect on mpg.

So, both h p statistically significant and weight is statistically significant.

(Refer Slide Time: 43:49)



Model 1 mpg=$\beta_0+\beta_1$wt+ϵ

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.285      1.878  19.858        0
wt             -5.344      0.559  -9.559        0
```

Model 2 mpg=$\beta_0+\beta_1$wt+$\beta_2$hp+ϵ

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.227      1.599  23.285    0.000
wt             -3.878      0.633  -6.129    0.000
hp             -0.032      0.009  -3.519    0.001
```

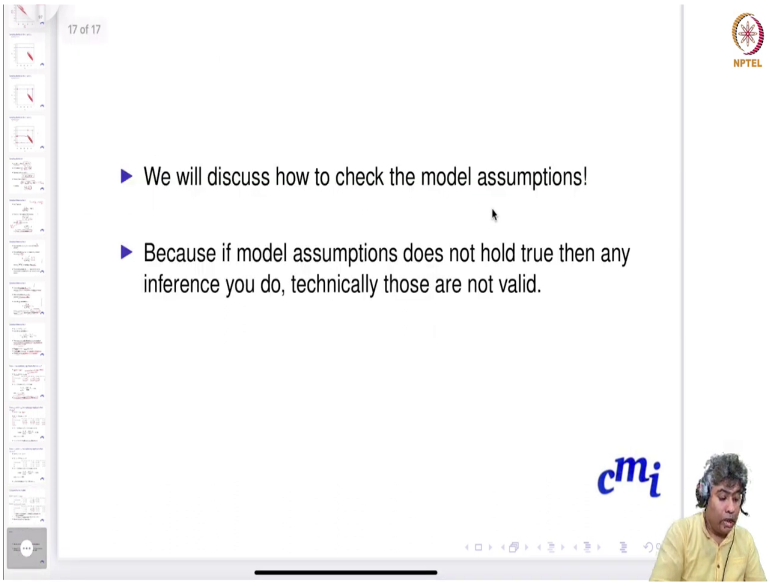
1. Model 1 is a 2D model, and Model 2 is a 3D model: Are they comparable?
2. The $se(\hat{\beta}_1)$ in Model 2 is higher than Model 1. Why?

- We will discuss these issues later.

So, what are the interesting thing? Now I am combining both the model ok. So, the first model is only with weight and miles per gallon and the second model is weight and horsepower with miles per gallon. So, the model 1 is a 2D model and model 2 is a 3D model are they comparable this is the first question we are going to ask. Standard error of beta 1 hat in model 2 is higher than the model 1.

So, you can see again here the standard error of beta 1 hat in the model 1 was 0.559 and it has gone up 0.633 why it has so? We will discuss these issues later these are very important issues and for now we are stopping here.

(Refer Slide Time: 44:47)



17 of 17

- We will discuss how to check the model assumptions!
- Because if model assumptions does not hold true then any inference you do, technically those are not valid.

So, we will discuss how to check the model assumptions because if model assumptions does not hold true then any inference you do technically those are not valid. So, we will stop here now we will do some hands on.

Thank you very much, see you in next video.