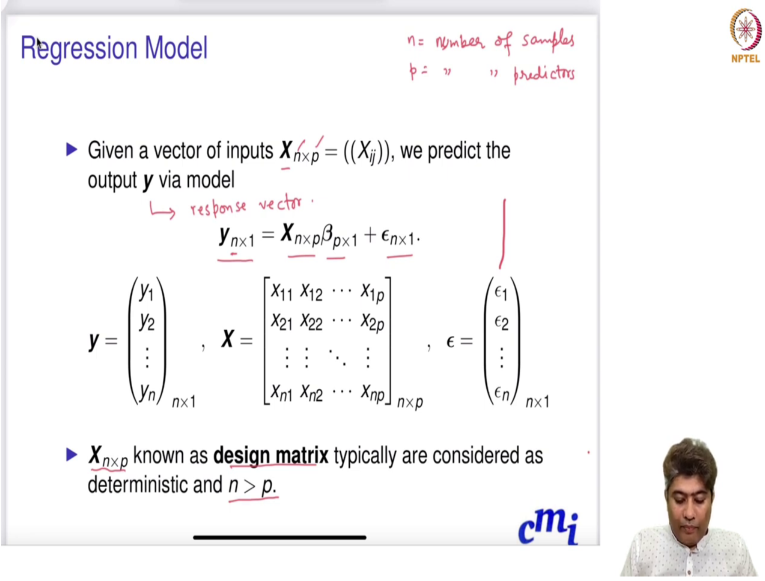
Predictive Analytics - Regression and Classification

Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 12
Gauss Markov Theorem

Hello all, welcome to the Predictive Analytics Regression and Classification course.

(Refer Slide Time: 00:27)



Regression Model

n= number of samples
p= " " predictors

- Given a vector of inputs $\underline{X}_{\underline{n\times p}} = ((X_{ij}))$, we predict the output y via model

↳ response vector.

$$\underline{y}_{n\times 1} = \underline{X}_{n\times p}\underline{\beta}_{p\times 1} + \underline{\epsilon}_{n\times 1}.$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

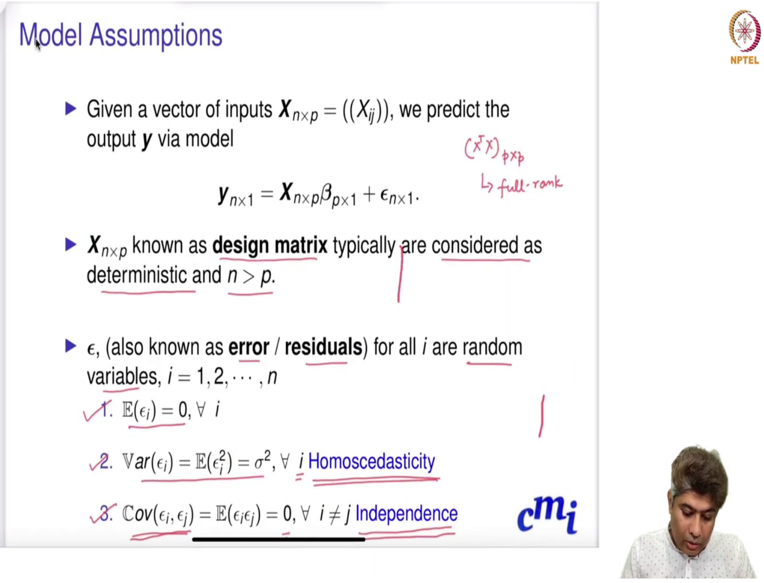
- $\underline{X}_{n\times p}$ known as **design matrix** typically are considered as deterministic and $\underline{n > p}$.

cmi

Today we are going to this is lecture 3 and part b today we are going to discuss some aspects of the Gauss Markov Theorem. Now, in the regression model we will start with the standard regression model where the vector of inputs is a design matrix of X of order n and p. Where, n is the n is the number of samples number of samples and p is number of predictors.

So, and y is the response vector response vector ok; so, this is our model y equal to X beta plus epsilon. So, with and our response vector y is a vector of size n additives and is a standard model. Now, now onwards we will consider X matrix is the is our design matrix and number of sample is greater than p, because if the number of sample is greater than p.

(Refer Slide Time: 02:06)



Model Assumptions

- Given a vector of inputs $\boldsymbol{X}_{n \times p} = ((X_{ij}))$, we predict the output $\boldsymbol{y}$ via model

$$\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

$(X^T X)_{p \times p}$

↳ full-rank

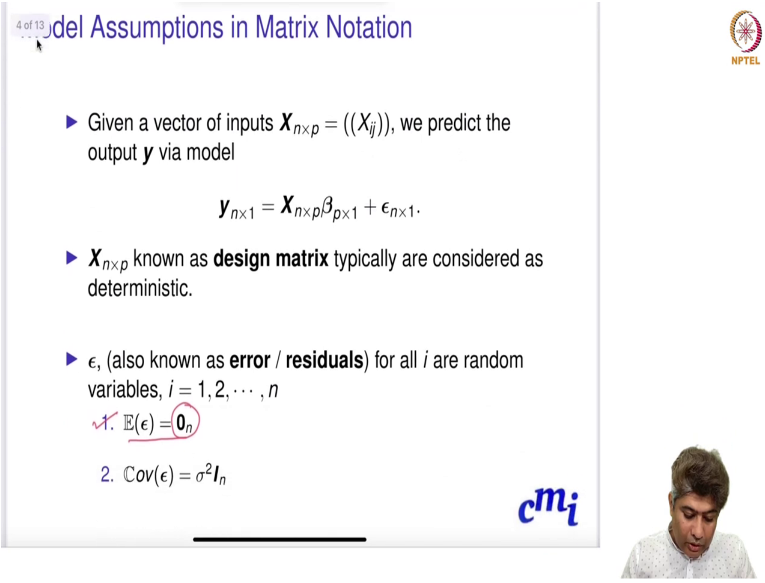
- $\boldsymbol{X}_{n\times p}$ known as **design matrix** typically are considered as deterministic and $n > p$.
- ϵ, (also known as **error** / **residuals**) for all i are random variables, $i = 1, 2, \cdots, n$
  1. ✓ $\mathbb{E}(\epsilon_i) = 0, \forall\ i$
  2. ✓ $\mathbb{V}ar(\epsilon_i) = \mathbb{E}(\epsilon_i^2) = \sigma^2, \forall\ i$ Homoscedasticity
  3. ✓ $\mathbb{C}ov(\epsilon_i, \epsilon_j) = \mathbb{E}(\epsilon_i \epsilon_j) = 0, \forall\ i \neq j$ Independence

Then X transpose X transpose X will be p cross p matrix and since it will be full rank matrix, X transpose X will be full rank matrix and we will have a unique solution. So, we will consider X design matrix typically considered as a deterministic and n is greater than p.

Now, epsilon also known as error or residual for all samples are considered as a random variables. And on these random variables we assume three assumptions; first expectation of epsilon i is equals to 0, second variance of epsilon i is sigma square for all i. For all samples that is the variance of residual is sigma square, since and this assumption typically known as Homoscedasticity.

And the third assumptions that we are making are known as covariance of epsilon i epsilon j equal to 0, this is coming from the independence assumption that is i th sample and the j th sample they are independent of each other. And that on that assumption gives us the assumption that covariance of epsilon i and epsilon j 0.

(Refer Slide Time: 04:01)



4 of 13 del Assumptions in Matrix Notation

- Given a vector of inputs $\boldsymbol{X}_{n\times p} = ((X_{ij}))$, we predict the output $\boldsymbol{y}$ via model

$$\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

- $\boldsymbol{X}_{n\times p}$ known as **design matrix** typically are considered as deterministic.

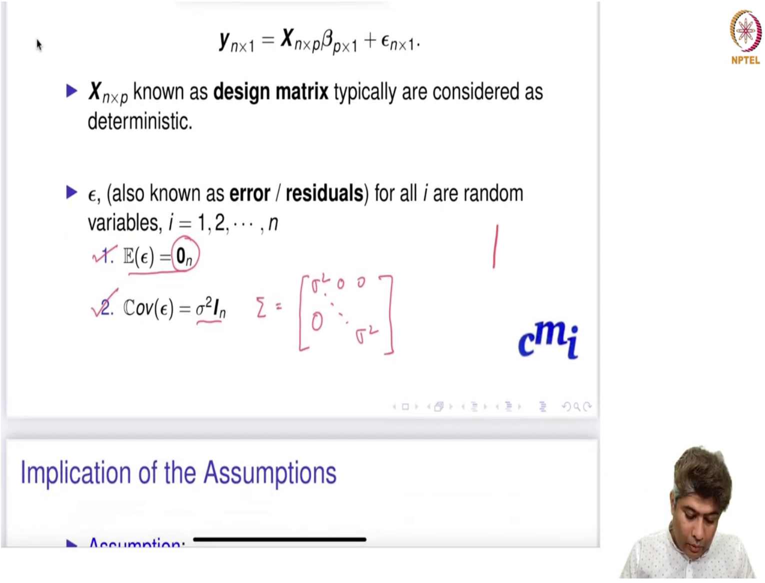
- ϵ, (also known as **error** / **residuals**) for all i are random variables, $i = 1, 2, \cdots, n$

  ✓1. $\mathbb{E}(\epsilon) = \mathbf{0}_n$

  2. $\mathbb{C}ov(\epsilon) = \sigma^2 \boldsymbol{I}_n$

Now; that means, this we can write it in a matrix notation where expectation of epsilon is 0 a vector of 0, n vector of size n where all elements are 0. And then we can join the second assumption and the third assumption a little bit, we can join this second and third assumption. And then we can make this assumption as with the matrix notation it will be covariance of epsilon is sigma square by i n, so; that means, it is a diagonal matrix whose.

(Refer Slide Time: 04:59)



$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

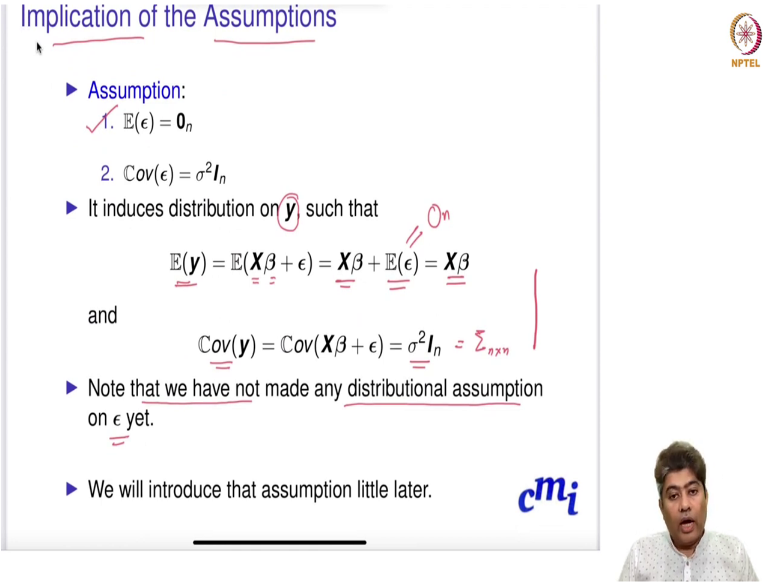
- $X_{n\times p}$ known as **design matrix** typically are considered as deterministic.
- ϵ, (also known as **error** / **residuals**) for all i are random variables, $i = 1, 2, \cdots, n$
  1. ✓ $\mathbb{E}(\epsilon) = \boxed{\mathbf{0}_n}$
  2. ✓ $\mathbb{C}ov(\epsilon) = \underline{\sigma^2 I_n}$ $\quad \Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \ddots & \\ & & \sigma^2 \end{bmatrix}$

Implication of the Assumptions

- Assumption:

So, basically capital sigma will be all sigma square all the diagonal elements are sigma square and all the off diagonal elements are all 0 ok.

(Refer Slide Time: 05:18)



Implication of the Assumptions

- Assumption:
  1. $\mathbb{E}(\epsilon) = \mathbf{0}_n$
  2. $\mathbb{C}ov(\epsilon) = \sigma^2 \mathbf{I}_n$
- It induces distribution on $\mathbf{y}$ such that

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{X}\beta + \epsilon) = \mathbf{X}\beta + \overset{0_n}{\mathbb{E}(\epsilon)} = \mathbf{X}\beta$$

and

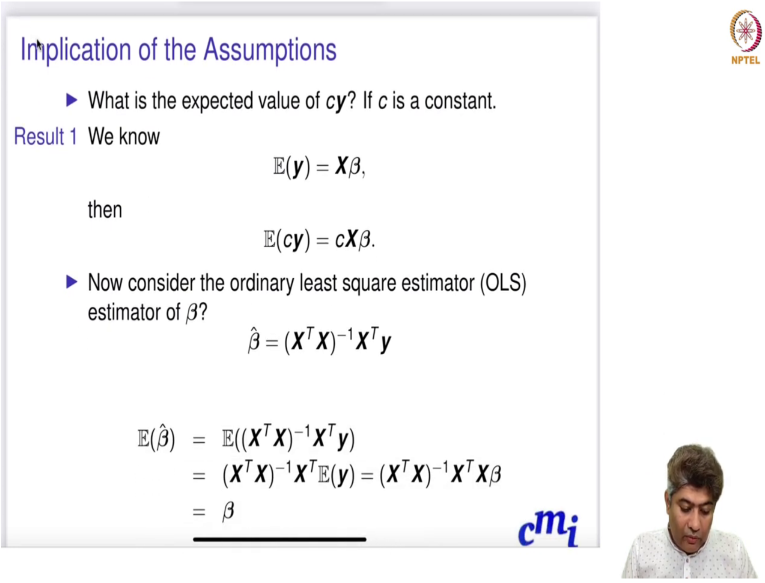
$$\mathbb{C}ov(\mathbf{y}) = \mathbb{C}ov(\mathbf{X}\beta + \epsilon) = \sigma^2 \mathbf{I}_n = \Sigma_{n \times n}$$

- Note that we have not made any distributional assumption on ϵ yet.
- We will introduce that assumption little later.

Now, what is it mean, what is the implication of this assumption, what is the implication of this assumption? The implication means, it induces a distribution on the response vector y. Because expectation of y is equal to we can say if we take expectation of y X typically we are considering a deterministic data in a classical setup we consider as a unknown constant.

So; that means, X beta plus expectation of epsilon, expectation of epsilon is 0 that is our under assumption this is the assumption under this assumption. So; that means, expectation of y is equal to X beta, and the covariance of y is equal to sigma square by i n. So, this is our covariance matrix of order n cross n typically diagonal matrix with all the off diagonal are 0.

(Refer Slide Time: 06:26)



Implication of the Assumptions

- What is the expected value of cy? If c is a constant.

Result 1 We know

$$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\beta,$$

then

$$\mathbb{E}(c\boldsymbol{y}) = c\boldsymbol{X}\beta.$$

- Now consider the ordinary least square estimator (OLS) estimator of β?

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

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}(\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta \\ &= \beta\end{aligned}$$

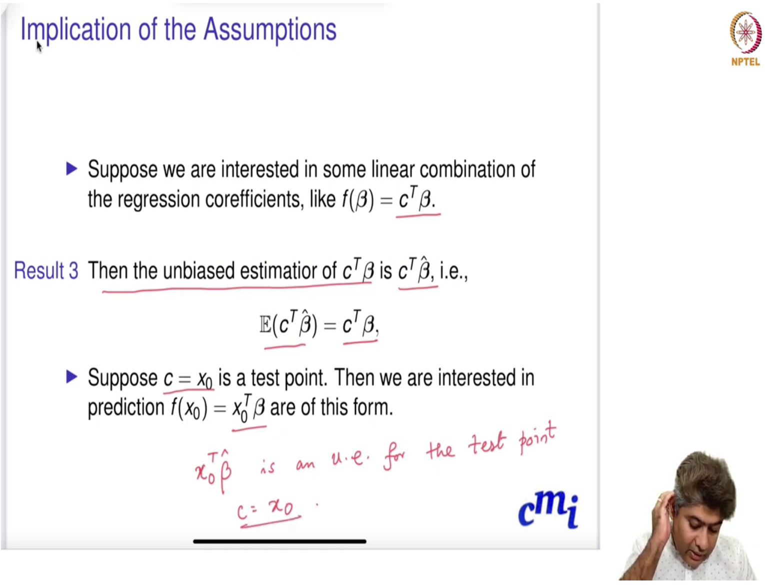
Now, one important thing I would like to make that we have not made any distributional assumption on epsilon yet. We have only assumed that the residual or errors are random variable; that means, they have a distribution they do have a distribution. But they do not and they have a mean vector, they have a covariance vector, but they we are not making any assumption about whether epsilon follow a multivariate normal or any other distribution.

We will introduce this assumption later, we will introduce assumption later as we require ok. Now, what is the implication we will continue on this, what is the implication of considering this assumption, what is the expected value of c y, if c is a constant? Now, if you we have already found expected y is equal to X beta, then expected value of c y will be expected value of c into X beta. This is an interesting thing, because now consider ordinary least square estimator ordinary least square estimator OLS estimator of beta.

So, this is our OLS estimator of beta correct, this is our OLS estimator of beta. Now, what is if we if you see X transpose X inverse X transpose this part is deterministic part like. Tut y is the non deterministic, but it is a response and then; that means, we can we take expected value of beta hat. And then in beta hat we replace the X transpose the formula the analytic form X transpose X inverse X transpose y and then that gives us X transpose X inverse X transpose a constant; so, it will come out of the expectation.

So, what we have is here is expectation of y and then expectation y is X beta. Now, interestingly this is X transpose X inverse X transpose X which will be identity matrix; so, we left with beta. So, what does it mean that OLS estimator beta hat is an unbiased estimator of beta. So, this is a OLS estimator is unbiased estimator of beta, this is a very important finding.

(Refer Slide Time: 09:41)



Implication of the Assumptions

- Suppose we are interested in some linear combination of the regression corefficients, like $f(\beta) = \underline{c^T\beta}$.

Result 3 $\underline{\text{Then the unbiased estimator of } c^T\beta}$ is $\underline{c^T\hat{\beta}}$, i.e.,

$$\mathbb{E}(\underline{c^T\hat{\beta}}) = c^T\underline{\beta},$$

- Suppose $\underline{c = x_0}$ is a test point. Then we are interested in prediction $f(x_0) = \underline{x_0^T}\beta$ are of this form.

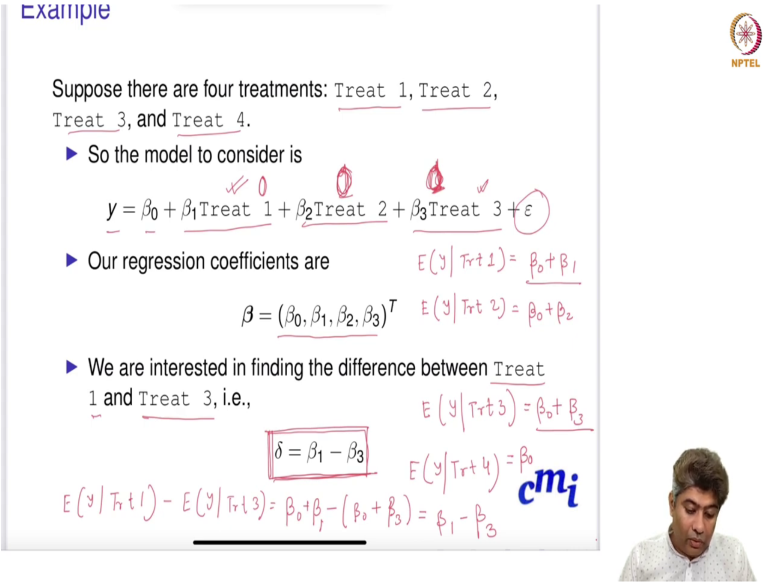
$x_0^T\hat{\beta}$ is an u.e. for the test point

$\underline{c = x_0}$.

Now, suppose we are interested in some linear combination of the regression coefficient; say some c transpose beta, what is the interpretation of this, we will come to that we will do an example and then it will become much more clear to you. So, the result says then unbiased estimator for c transpose beta is simply c transpose beta hat, because effectively expected value of c transpose beta hat is if you just take c transpose out of the expectation then it will be like c transpose of beta hand which is beta; so, c transpose beta hat.

So, this is right away c transpose beta hat is an unbiased estimator. Now, suppose c is equal to x naught a test point, then we are interested in finding f of x transpose beta, x transpose x naught transpose beta that is my target. So, all I have to do that x naught transpose beta hat is my unbiased estimator is an unbiased estimator for the test point test point c equal to x naught; so, this is an interesting findings that we got.

(Refer Slide Time: 11:34)



Example

Suppose there are four treatments: Treat 1, Treat 2, Treat 3, and Treat 4.

- So the model to consider is

$$y = \beta_0 + \beta_1 \underset{}{\overset{0}{\texttt{Treat 1}}} + \beta_2 \overset{0}{\texttt{Treat 2}} + \beta_3 \overset{0}{\texttt{Treat 3}} + \varepsilon$$

- Our regression coefficients are

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$$

- We are interested in finding the difference between Treat 1 and Treat 3, i.e.,

$$\boxed{\delta = \beta_1 - \beta_3}$$

$E(y \mid Trt\,1) = \beta_0 + \beta_1$

$E(y \mid Trt\,2) = \beta_0 + \beta_2$

$E(y \mid Trt\,3) = \beta_0 + \beta_3$

$E(y \mid Trt\,4) = \beta_0$

$E(y \mid Trt\,1) - E(y \mid Trt\,3) = \beta_0 + \beta_1 - (\beta_0 + \beta_3) = \beta_1 - \beta_3$

Now, another example we will do, that suppose there are four treatments; treatment 1, treatment 2 treatment 3 and treatment 4, these are the four treatment. And we consider the model y equal to beta naught plus beta 1 times treatment 1, beta 2 times treatment 2, beta 3 times treatment 3 and an epsilon ok. Now, our regression coefficient beta equal to beta naught plus beta 1 beta 3 beta 3, this is the regression coefficient.

Now, what is treatment 1? So, what we have done as usual as discussed last time, this is 1 hit 1 hot encoding. So, if I need to figure out what would be the expected value of y when it is treatment 1 what I will do? I will put in this case treatment equal to 1 and here I will put zeros. Now, if I need a to calculate what would be the expected value of y if I am looking for treatment 2, then what I will do? I will make this one as 0 and for treatment 2, I will make treatment 2 as treatment 1 and treatment 3 will be 0.

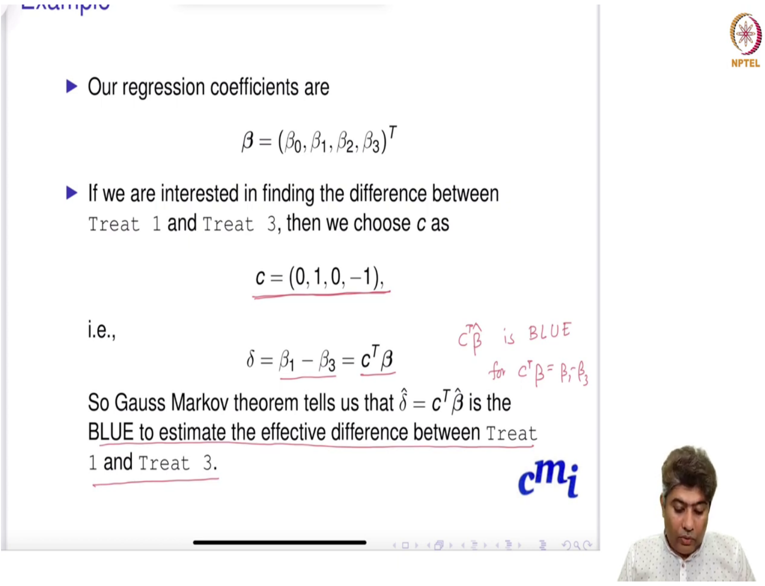
So, the model will be; so, the solution will be beta naught plus beta 2, and in this case it will be beta naught plus beta 1. Now, if you are interested in expected value of y, when if it is treatment 3 if it is treatment 3, then same way I will take treatment 3 as 1 and take treatment 1 as 1 and 2 as 0. And I will make treatment 3 as 1 and that will give me beta naught plus beta 3.

Now, if I am interested in beta expected value of y when it is treatment 4, then what I have to do I will put all treatment 1 equal to 0, treatment 2 equal to 0, treatment 3 equal to 0. And intercept on beta naught is itself will give me the expected value or expected level of y if we are in treatment 4. Now, suppose we are interested in the to measure the difference between treatment 1 and treatment 3; so, this is what we are interested in, what is the difference between beta 1 and beta 3? Beta 1 is the measure of difference one.

So, basically this is what we want to do expected value of y given treatment 1 minus expected value of y given treatment 3. Now, what is y given in treatment 1 y given treatment 1 is beta naught plus beta 1 beta naught plus beta 1 minus what is y given treatment 3, this is beta naught plus beta 3 beta naught plus beta 3; so, what we left with beta one minus beta 2.

So, if I am interested in measuring the difference between the expected level for treatment 1 and expected level for treatment 3, then basically the difference between beta 1 and beta 3 is good enough. If I just measure the difference between beta 1 and beta 3 that would be good enough.

(Refer Slide Time: 16:08)



Example

- Our regression coefficients are

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$$

- If we are interested in finding the difference between `Treat 1` and `Treat 3`, then we choose c as

$$c = (0, 1, 0, -1),$$

i.e.,

$$\delta = \beta_1 - \beta_3 = c^T\beta$$

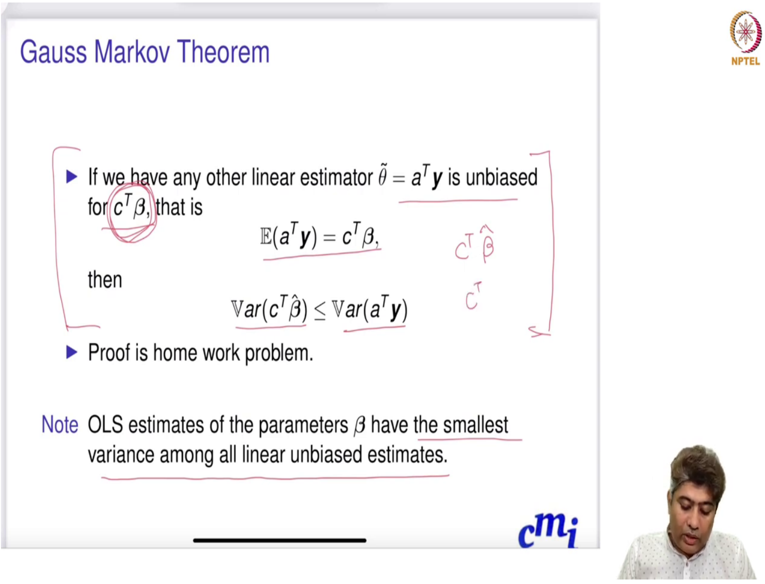
$c^T\hat{\beta}$ is BLUE for $c^T\beta = \beta_1 - \beta_3$

So Gauss Markov theorem tells us that $\hat{\delta} = c^T\hat{\beta}$ is the BLUE to estimate the effective difference between `Treat 1` and `Treat 3`.

So, for that what we will do we will simply consider c to be 0 1 0 minus 1 and you see if you take c transpose beta and then what you will get? Beta 1 minus beta 3. Now, gauss Markov theorem tells me that if I take c transpose beta hat just take the beta OLS estimator beta hat multiply with this c or the take the dot product of the c with the c and the beta hat.

That will be best linear unbiased estimator for the effective difference between treatment 1 and treatment 3. So, this will be c transpose beta hat is based linear unbiased estimator for c transpose beta which is beta 1 minus beta 3 in my case. So, there are these kind of very nice things you can do for using gauss Markov theorem.

(Refer Slide Time: 17:24)



Gauss Markov Theorem

- If we have any other linear estimator $\tilde{\theta} = a^T \mathbf{y}$ is unbiased for $c^T\beta$, that is
$$\mathbb{E}(a^T\mathbf{y}) = c^T\beta,$$
then
$$\mathbb{V}ar(c^T\hat{\beta}) \leq \mathbb{V}ar(a^T\mathbf{y})$$
- Proof is home work problem.

$c^T\hat{\beta}$

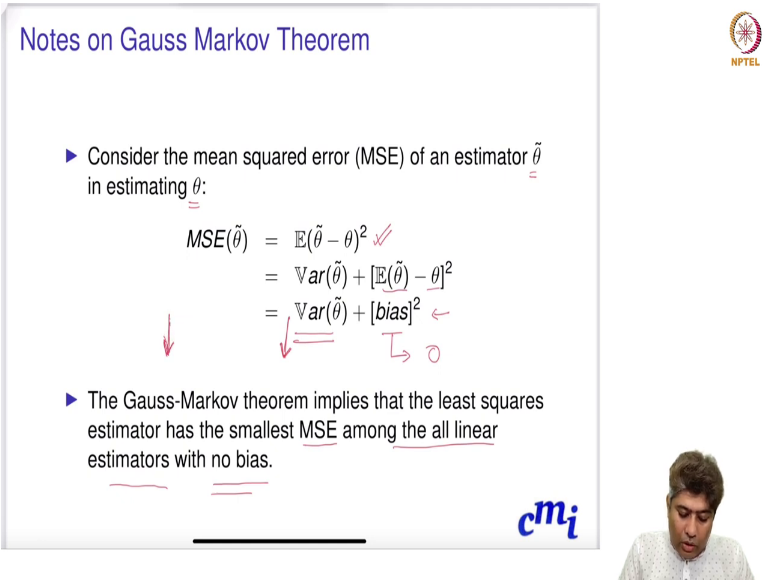
c^T

Note OLS estimates of the parameters β have the smallest variance among all linear unbiased estimates.

Now, if you have any other linear estimator say theta tilde a transpose y is unbiased estimator for c transpose beta. Then you can show expected value of a transpose y is equal to c transpose beta and variance of c transpose beta hat is less than equal to variance of a transpose y. So, this is your, this is what typically called the gauss Markov theorem. So; that means, essentially OLS estimator estimate of the parameter beta is in the smallest variance have the smallest variance among the linear unbiased estimates.

So, this is the main crux of the gauss Markov theorem and that is that makes our life very simple. All you have to do that if I need a good; so, if I am interested in finding an estimator for c transpose beta which is turns out to be very useful form functional form like. And then what all I have to do? I have to just find a OLS estimator and then multiply the OLS estimator beta hat with c transpose. That will be that will ensure that this will be base linear unbiased estimator for my target functional form c transpose beta.

(Refer Slide Time: 19:11)



Notes on Gauss Markov Theorem

- Consider the mean squared error (MSE) of an estimator $\tilde{\theta}$ in estimating θ:

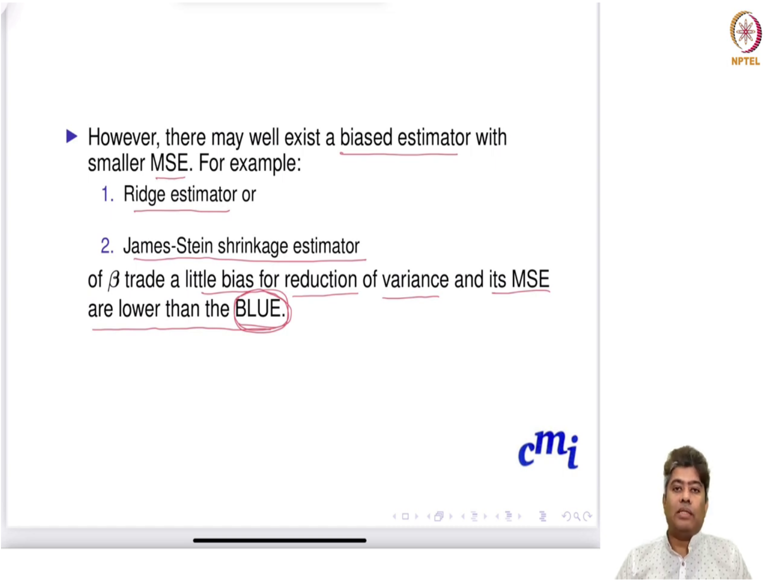
$$
\begin{aligned}
MSE(\tilde{\theta}) &= \mathbb{E}(\tilde{\theta} - \theta)^2 \\
&= \mathbb{V}ar(\tilde{\theta}) + [\mathbb{E}(\tilde{\theta}) - \theta]^2 \\
&= \mathbb{V}ar(\tilde{\theta}) + [bias]^2
\end{aligned}
$$

- The Gauss-Markov theorem implies that the least squares estimator has the smallest MSE among the all linear estimators with no bias.

Now, what is it there are certain things to note here, if you consider mean square error for an estimated theta tilde. Suppose, theta is your target parameter and theta tilde is your estimator that you are using to estimate. Now, mean square error is simply expected value of theta tilde minus theta whole square. So, you can write this as variance of theta tilde plus expected value of theta tilde minus theta whole square which is a bias; now, so, basically variance plus bias square.

Now, gauss Markov theorem implies that least square estimated has smallest mean square error among the all linear estimator with no bias; now, bias here is 0. So, among all the unbiased estimator the least square estimator and this c transpose beta hat has the least mean square error.

(Refer Slide Time: 20:36)



- However, there may well exist a biased estimator with smaller MSE. For example:
  1. Ridge estimator or
  2. James-Stein shrinkage estimator

  of β trade a little bias for reduction of variance and its MSE are lower than the BLUE.

However, there may well exist a biased estimator with smaller MSE; for example, ridge estimator or James-Stein shrinkage estimator of beta trade a little bias for reduced reduction of variance and its MSE are lower than the base linear unbiased estimator. So, you can find the estimator though gauss Markov theorem is very good strong theorem. It shows that among the class of all linear unbiased estimator; it has the minimum mean square error, it has the minimum variance.

But estimator like ridge estimator James-Stein estimator; you can trade a little bit bias you can induce a little bit bias, but you will if you can reduce variance great significantly. Then overall mean square error can go significantly reduce and in that case what will happen that mean square error is even lower than the base linear mean square error of the base linear unbiased estimator.

So, that so, sometimes you can come up with a slightly biased estimator which will be; which will have a better mean square in a lower mean square error than the gauss Markov estimator. However, among the class of all unbiased estimator gauss square gauss Markov estimator is the best; so, we will stop now.

Thank you very much lets join to the next video.