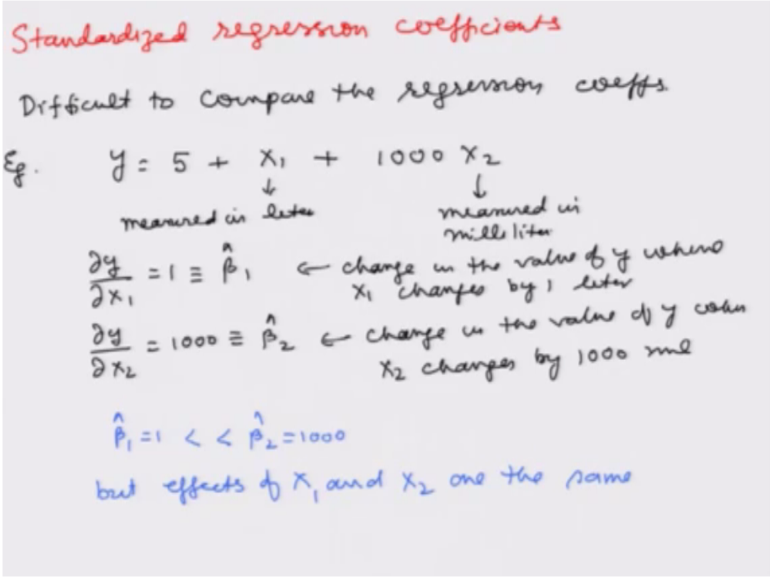
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture – 14
Standardized Regression Coefficients and Testing of Hypothesis

(Refer Slide Time: 00:20)



Standardized regression coefficients

Difficult to compare the regression coeffs.

Eg. $y = 5 + \underset{\substack{\downarrow \\ \text{measured in liter}}}{x_1} + 1000\underset{\substack{\downarrow \\ \text{measured in milli liter}}}{x_2}$

$\frac{\partial y}{\partial x_1} = 1 \equiv \hat{\beta}_1$ ← change in the value of y when x_1 changes by 1 liter

$\frac{\partial y}{\partial x_2} = 1000 \equiv \hat{\beta}_2$ ← change in the value of y when x_2 changes by 1000 ml

$\hat{\beta}_1 = 1 << \hat{\beta}_2 = 1000$

but effects of x_1 and x_2 are the same

Welcome to the lecture, now in this lecture, we are first going to talk about standardized regression coefficients. You see when we are trying to fit a multiple linear regression model then it is based on different types of independent variables like as x1, x2, x k and these variables are measured in different units. For example, x 1 can be in kilogram, x 2 can be in liter, x 3 can be in some other unit.

So, when we are trying to compute the regression coefficients, they are trying to denote the rate of change in the value of y that is the output when there is a unit change in the corresponding value of independent variable. Once these variables are measured in different units, sometimes it becomes difficult to compare the regression coefficients. So in such a situation, it is difficult to compare the regression coefficients.

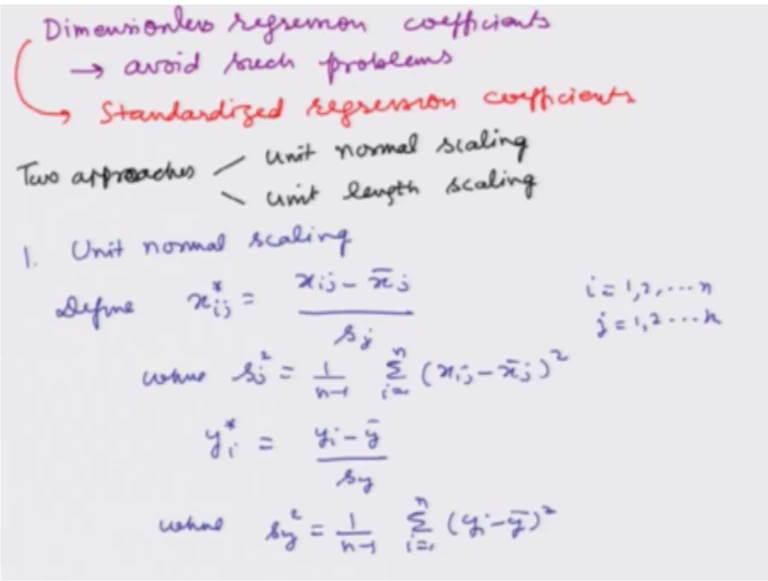
If, I try to take an example, suppose the fitted model is y = 5 + say x1, + 1000 times x 2, and suppose, this x 1 is a variable, which is measured in see here liter and x 2 is a variable, which is measured in milliliter, so now if you try to see partial derivative of y with respect to x 1, it

= here one and that is essentially your beta1 hat and partial derivative of y with respect to x 2 is here 1000, which is essentially the value of beta hat2.

So, if you try to see here, the value of beta1 hat this is denoting the change in the value of y when x1 changes by one liter and this is indicating the change in the value of y when x two changes by 1000 milliliters. Now, if you try to observe the value of beta1 hat, which is here = 1, this is much, much smaller than the value of the beta 2 hat, which = here 1000.

But if you try to see both of them are denoting the same change, right, this is in liter and whereas this is in milliliter and one liter = 1000 milliliter. so, if you try to see the values of beta1 hat and beta2 hat are indicating that x1 and x2 have got different effects, but effects of x1 and x2 are the same. So now in such cases, we have a problem that how to compare the different regression coefficient.

(Refer Slide Time: 04:23)



Dimensionless regression coefficients

→ avoid such problems

→ Standardized regression coefficients

Two approaches ↗ unit normal scaling ↘ unit length scaling

1. Unit normal scaling

Define $x_{ij}^* = \dfrac{x_{ij} - \bar{x}_j}{s_j}$ $i = 1, 2, \cdots n$, $j = 1, 2 \cdots k$

where $s_j^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$

$y_i^* = \dfrac{y_i - \bar{y}}{s_y}$

where $s_y^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

And in such situation, one option is that we can work with dimensionless regression coefficients and the use of dimensionless regression coefficient will avoid such problems. Now, the next question is how to obtain the dimensionless regression coefficient and this dimensionless regression coefficients, they are actually called as standardized regression coefficients.

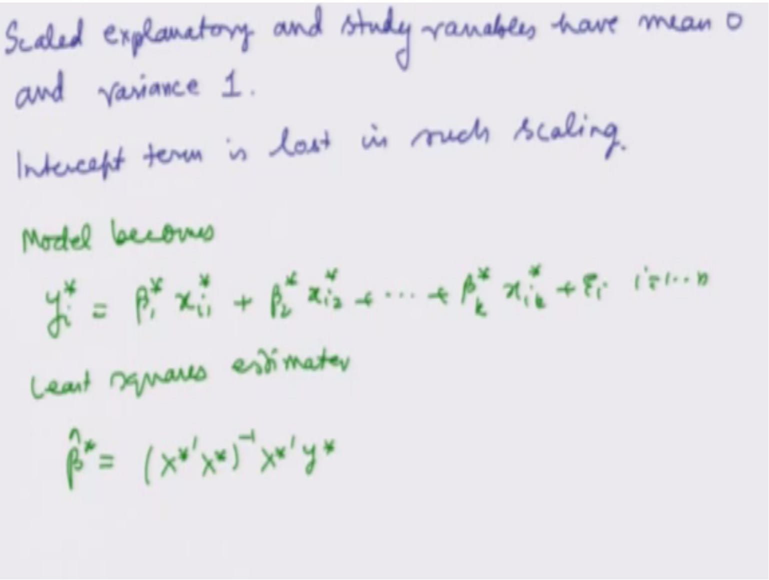
So, in order to obtain such standardized regression coefficients, we have two approaches, one is unit normal scaling procedure and another is unit length scaling procedure. In both the procedures, we try to change the values of study and explanatory variable, so we try to discuss them one by one. So, first of all I try to discuss the unit normal scaling procedure. In this case, we try to define the values of explanatory variables.

So let us denote by say xij star as the original observation xij minus the sample mean of the observation on the corresponding explanatory variable divided by sj, where this sj square is given by one over n - 1, summation i goes from 1 to n xij - x bar j whole square. So, we are trying to take a particular explanatory variable, we try to collect the observation on them, we try to find out their mean and their sample variance.

And we try to subtract every observation xij by its corresponding mean and divided by the corresponding standard deviation, and similarly we try to transform y as yi star and in this case also we try to subtract the original observations y i by their sample means y bar and divided by their standard deviation where s y square is 1 over n - 1, summation i goes from 1 to n y i - y bar whole square.

And in this case if you recall, we had i goes from one to here and n and j goes from one to k because we have k explanatory variables.

(Refer Slide Time: 07:36)



Scaled explanatory and study variables have mean 0 and variance 1.

Intercept term is lost in such scaling.

Model becomes

$$y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \cdots + \beta_k^* x_{ik}^* + \varepsilon_i \quad i=1\ldots n$$

Least squares estimator

$$\hat{\beta}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$$

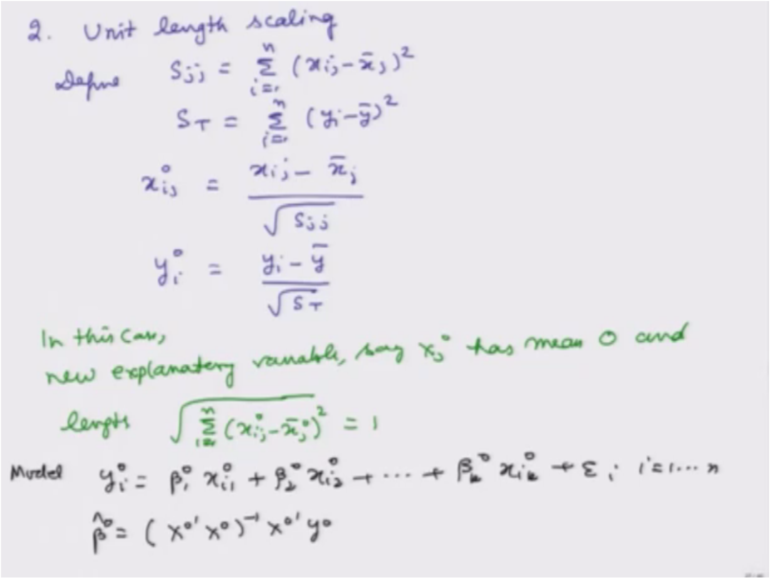
So, when we try to scale the explanatory and the steady variables, then what happens that this scaled explanatory and steady variables have mean 0 and variance unity that is one, and when we are doing such scaling, then intercept term is lost in such a scaling. Now, I can transform my original model, which was based on y and x1 x2 x k in terms of the scaled variables y star and x star.

So our model becomes say y i star = beta1 star x i one star + beta2 star x i 2 star up to here say beta k star x i k star, plus epsilon i, i goes from 1 to n, and based on that, we can obtain the least square estimator or even equivalently the maximum likelihood estimator also under the assumption that epsilons are following the normal distribution with mean 0 and variance sigma s square and their IID beta hat, let us denote it by beta hat star.

And this becomes here x star transpose x star, whole inverse x star transpose y star, and, where this x star is the matrix, which is obtained by the scale observations on x1 x2 x k and y star is a vector of observations on the steady variable, which are obtained after the scaling.

In this case, if you try to see, we are trying to standardize the observation like in the same way as we do in the case of normalizing the random variable that is a random variable - its mean divided by the standard deviation or a standard error. So that is why this procedure is called as unit normal scaling procedure.

(Refer Slide Time: 10:15)



2. Unit length scaling

Define $S_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$

$S_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$$x_{ij}^{\circ} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}$$

$$y_i^{\circ} = \frac{y_i - \bar{y}}{\sqrt{S_T}}$$

In this case, new explanatory variable, say X_j° has mean 0 and length $\sqrt{\sum_{i=1}^{n} (x_{ij}^{\circ} - \bar{x}_j^{\circ})^2} = 1$

Model $y_i^{\circ} = \beta_1^{\circ} x_{i1}^{\circ} + \beta_2^{\circ} x_{i2}^{\circ} + \cdots + \beta_k^{\circ} x_{ik}^{\circ} + \varepsilon_i \quad i = 1 \ldots n$

$\hat{\beta}^{\circ} = (X^{\circ\prime} X^{\circ})^{-1} X^{\circ\prime} y^{\circ}$

So now we discuss the next procedure, which is unit length scaling. In this case, we define two quantities say sjj, which = i goes from 1 to n xij - x bar j whole square and say ST, which = i goes from 1 to here n y i- y bar whole square. So if you observe these quantities are similar to the earlier procedure of unit normal scaling, but sjj and ST, they are based on the quantity similar to in the analysis of variance that we are going to discuss later on.

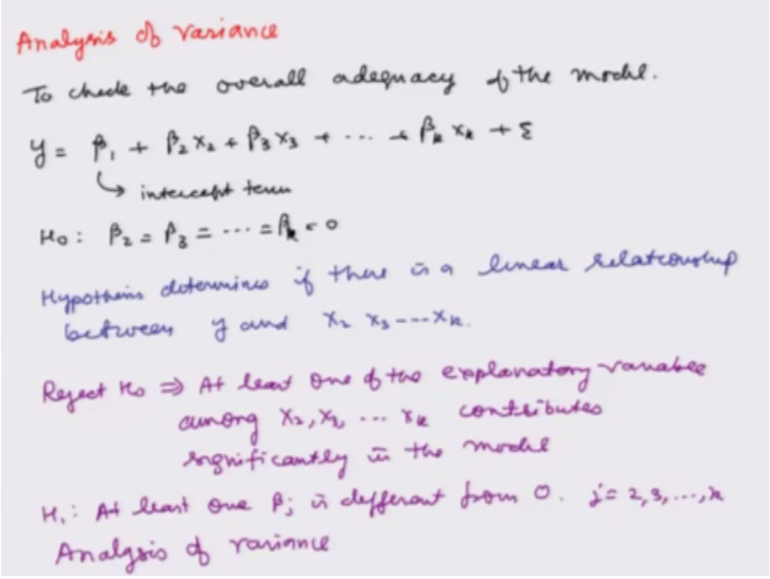
And, based on that sjj and ST, we try to now scale our explanatory variables and steady variable as xij, let me now denote it by naught. So this will be the original observations xij on the jth explanatory variable minus the sample mean of the jth explanatory variable and square root of sjj, and similarly we try to scale the observation on steady variables as an original observation, y i - its mean divided by square root of ST.

In such a case what happened that the new explanatory variable say x j naught has mean 0 and length, which is obtained by square root of i goes from 1 to n, xij naught - x bar j naught whole square, this = here one. So that is why this procedure is called as unit length scaling, and in this case, now the model can be rewritten in terms of scaled variables as y i naught = beta1 naught x i 1 naught + beta 2 naught x i 2 naught + beta k naught x i k naught + epsilon i, i goes from 1 to n.

And then in case we try to find out the estimator of beta as beta hat, which can be the least square estimator or the maximum likelihood estimator assuming the normal quick distribution for epsilon i's this turns out to be x naught transpose, x naught whole inverse x naught transpose y naught, where x naught is the matrix of observations on k explanatory variables, after they are standardized.

And similarly y naught is the vector of observations on a steady variable after they've been say scaled.

(Refer Slide Time: 13:50)



Analysis of Variance

To check the overall adequacy of the model.

$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$

↳ intercept term

$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$

Hypothesis determines if there is a linear relationship between y and $x_2\ x_3 \cdots x_k$.

Reject $H_0 \Rightarrow$ At least one of the explanatory variables among $x_2, x_3, \cdots x_k$ contributes significantly in the model

H_1: At least one β_j is different from 0. $j = 2, 3, \ldots, k$

Analysis of variance

Next, we come on the aspect of now test of hypothesis. So, we are going to first discuss here test of hypothesis, which is called as analysis of variance, right. This is a test which is used to check the overall adequacy of the model. Next question is what do we really understand by the overall adequacy? You see a model is obtained by estimating the model parameters beta1, beta2, beta k, and we want to check here whether all the variables are significant or not.

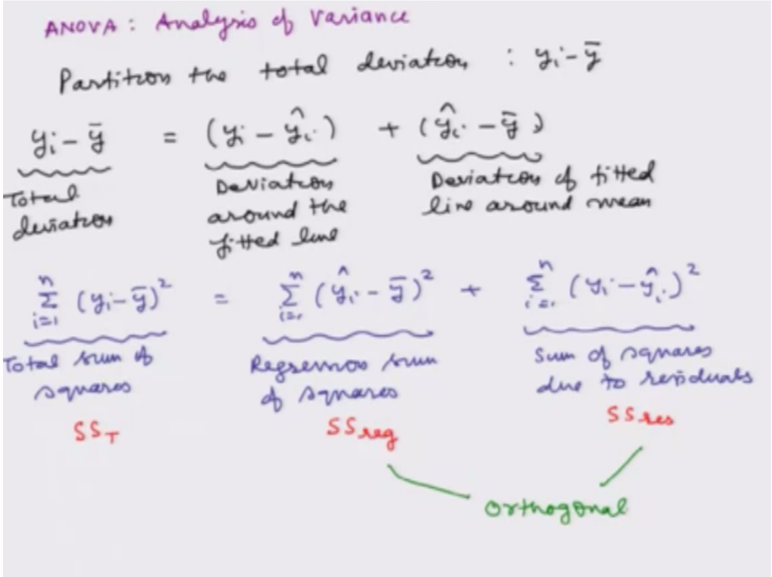
So, we consider here a model say beta1 + beta2 x 2 + beta3 x three up to here beta k x k + epsilon, and here we assume that all the observations on explanatory variable x1 that takes the value1, so that beta1 here is the intercept term, and all beta2 beta3 betak, they are the slope parameter. We are intentionally considering the presence of intercept term in this model because that we are going to relate with coefficient of determination that we are going to discuss later on

So this test of hypothesis is about testing whether all beta2 beta3 up to beta k they are 0 or not. So obviously if you try to see here, if all beta2 beta3 beta k, they become 0 that means my model becomes only here y = beta1 + epsilon. So this is a casting the significance of all the slope parameters together, and this hypothesis determines, if there is a linear relationship between y and all x2 x3 see here x k.

Now there are two options, whether this h naught is accepted or h naught is rejected. Well, obviously if h naught is accepted that means all the regression coefficient have got value 0, so all the corresponding explanatory variable x2, x3, x k they are not contributing in explaining the variation in y, but the rejection of h naught implies that at least one of the explanatory variable among x2, x3, x k contributes significantly in the model.

And in this case our alternative hypothesis is that at least one beta j is different from 0, where j goes from two, three up to k and the test procedure to test such a hypothesis, this is called as analysis of variance. So let us now try to briefly discuss that how this test of analysis of variance is obtained, and this analysis of variance shortly called as ANOVA.

(Refer Slide Time: 18:22)



ANOVA : Analysis of Variance

Partition the total deviation : $y_i - \bar{y}$

$$\underbrace{y_i - \bar{y}}_{\text{Total deviation}} = \underbrace{(y_i - \hat{y}_i)}_{\text{Deviation around the fitted line}} + \underbrace{(\hat{y}_i - \bar{y})}_{\text{Deviation of fitted line around mean}}$$

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\substack{\text{Total sum of squares} \\ SS_T}} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\substack{\text{Regression sum of squares} \\ SS_{reg}}} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\substack{\text{Sum of squares due to residuals} \\ SS_{res}}}$$

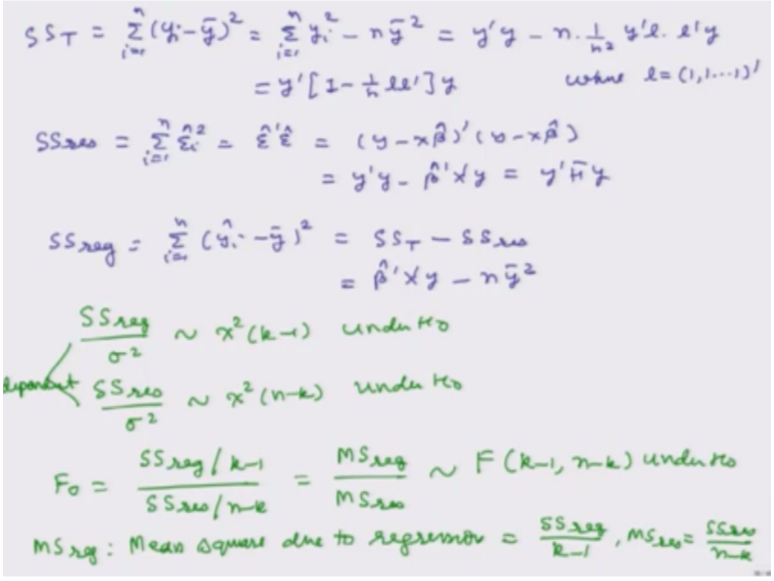
This ANOVA technique is actually based on partitioning the total variation, in this case we try to partition the total deviation, which is actually here y i minus y bar, and we try to divide it into 2components, for example I can write y i minus y bar = y i - y i hat + y i hat - y bar. So if you observe here, this quantity that is giving us total deviation and this is a deviation around the fitted line and this is a deviation of fitted line around mean that is y bar.

So now what we do, we try to square on both the sides and we sum them up from i goes from 1 to n. So when I square it, this becomes y i - y bar whole square and when I try to sum, this becomes i goes from 1to n, and when I try to square on the right hand side, I get the square quantities of both these factors + their cross product term. The cross product terms term becomes 0, so we get here summation i goes from one to n y i hat - y bar whole square + summation i goes from 1 to n y i - y hat whole square.

Now this quantity is called as total sum of squares and this quantity is called as regression sum of squares and this quantity is called as sum of squares due to residuals, and this is denoted as SST and this is denoted as SS reg and this is denoted as SS res that is our earlier notation also and one important thing we have to note it down here that when we try to partition the total sum of squares SST into two components sum of squares due to regression and some of square due to residual then both these terms are orthogonal to each other.

So this analysis of variance techniques is based on the partitioning the total variation into two different orthogonal components sum of squares due to regression and some of square due to residuals.

(Refer Slide Time: 22:23)



$$SS_T = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = y'y - n.\frac{1}{n^2}\, y'\ell.\ell'y$$

$$= y'[I - \tfrac{1}{n}\ell\ell']y \qquad \text{where } \ell = (1,1\ldots 1)'$$

$$SS_{res} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \hat{\varepsilon}'\hat{\varepsilon} = (y - X\hat{\beta})'(y - X\hat{\beta})$$

$$= y'y - \hat{\beta}'X y = y'\bar{H}y$$

$$SS_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SS_T - SS_{res}$$

$$= \hat{\beta}'Xy - n\bar{y}^2$$

$$\frac{SS_{reg}}{\sigma^2} \sim \chi^2(k-1) \text{ under } H_0$$

independent

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2(n-k) \text{ under } H_0$$

$$F_0 = \frac{SS_{reg}/k-1}{SS_{res}/n-k} = \frac{MS_{reg}}{MS_{res}} \sim F(k-1, n-k) \text{ under } H_0$$

MS_{reg} : Mean square due to regression $= \frac{SS_{reg}}{k-1}$, $MS_{res} = \frac{SS_{res}}{n-k}$

Now based on that, we can write down the expression more compactly before we go for the further analysis. Your SST here is nothing but i goes from 1 to n yi- y bar whole square this can be written as i goes from one to n y i square- n times y bar square and this can be further written as y transpose y - n into one over n square y transpose l into l transpose y where l is a column vector of all elements one.

And this can be further written as say here say y transposes i-1 over n ll prime times' y. So you can see that this quantity quite resembles with the metrics like i- x x transpose x whole inverse x transpose. Similarly, sum of square due to residual is nothing but, i goes from 1 to n epsilon i hat whole square and this I can write down as epsilon hat transpose epsilon hat and which can be written as y- x beta hat transpose y - x beta hat.

And when I try to open it this becomes y transpose y- beta hat transpose x transpose y, and this we had also written earlier as a y transpose h bar y and the sum of square due to regression, this is nothing but your i goes from one to n y i hat- y bar whole square and we note down that we also have established that total sum of square = sum of square due to residual and some of square due to regression.

So I can write down sum of square due to regression as SST- SS res and if I try to substitute all these values we get here beta hat transpose x transpose y- n times y bar square. So these are the three components, which are obtained by partitioning the total sum of squares into two orthogonal component sum of the square due to regression and some square due to residuals.

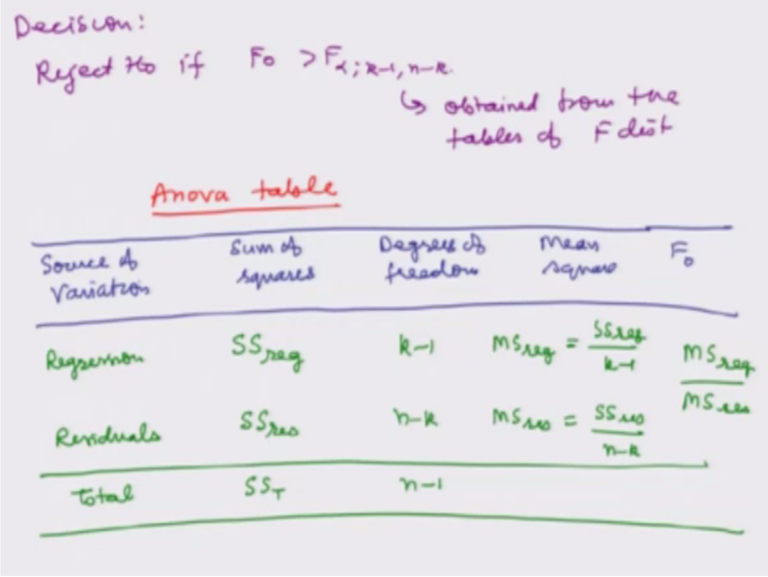
Well, I'm trying to briefly describe what are we going to do? This quantity sum of square due to regression divided by sigma square this follows a chi-square distribution with k- 1 degrees of freedom under h naught and SS residual divided by sigma square this follows a chi-square with n- k degrees of freedom under h naught and this concept we had discussed earlier also and both of them they are independent, right.

Now I have 2random variables, which are chi-squared distributed with a certain degrees of freedom, both are independent. So I can use here the F statistics and we define here statistics F naught say SS reg divided by its degrees of freedom upon SS res divided by its degrees of freedom and we try to call it say here mean square due to regression divided by mean square due to residual where your MS reg is the mean square due to regression.

This is actually means any mean square is divided by sum of its squares due to the aggression, in this case divided by the degrees of freedom, and similarly MS res this is also divided by SS res divided by degrees of freedom and - k and this will follow a F distribution with k - 1 and say n minus k degrees of freedom under h naught.

So now in simple words, if I want to test the hypothesis h naught beta two = beta3 = beta K. We simply have to compute sum of the square due to regression, sum of the square due to residual and then we have to obtain the statistics F naught.

(Refer Slide Time: 26:55)



Decision:

Reject H_0 if $F_0 > F_{\alpha; k-1, n-k}$

$\hookrightarrow$ obtained from the tables of F dist

Anova table

Source of Variation	Sum of squares	Degrees of freedom	Mean square	F_0
Regression	SS_{reg}	$k-1$	$MS_{reg} = \frac{SS_{reg}}{k-1}$	$\frac{MS_{reg}}{MS_{res}}$
Residuals	SS_{res}	$n-k$	$MS_{res} = \frac{SS_{res}}{n-k}$	
Total	SS_T	$n-1$		

Once I have obtained the statistics F naught, this follows F distribution with the degrees of freedom k minus 1 and n - k under H naught. So, now I can write down my decision, which is reject h naught, if f naught is greater than F alpha with degrees of freedom k- 1 and say n - k, where these values are obtained from the tables of F distribution. Now this entire procedure is expressed in the form of ANOVA table, which is actually analysis of variance table.

This analysis of variance table is constructed like follows and advantage of understanding this ANOVA table is that when we are using the software, the outcome is given in terms of analysis of variance table. So, it is important for us to understand that how the different ingredients of that ANOVA table have been computed and what are their interpretations.
So, this ANOVA table will have the first component as see here what is the source of variation?

The second component will be, what is the value of corresponding sum of squares, then it will also mention the degree of freedoms and then it will mention the mean square and then at the end it will give the value of F naught. So in this case we have the first source of variation, which is regression, due to regression. We are trying to fit a model, so the entire variability of the model is partitioned into two components.

One we are trying to capture through the fitted model, and that is controlled by sum of square due to regression and the path, which is beyond our control that is due to random variation that is being controlled by the sum of the square due to error or sum of square due to residuals. So the first component of variation is regression, second component is residuals and after that the third component is there total.

So, sum of square due to regression, we have obtained and we have denoted by SS reg, sum of square due to residual, this is obtained and we have denoted by SS res and total sum of squares is obtained and denoted as SST. The corresponding degrees of freedom are k- 1, n - k and their sum this is n- k + 1 k - 1 which = here n - 1.

And based on that we have obtained the mean square due to regression and mean square due to residual, which are obtained as SS reg divided by degrees of freedom, which is k - 1 and this is sum of square due to residual divided by the degrees of freedom and - k and then the value of F naught is obtained here by MS reg divided by MS res. So this is about the analysis of variance table that is obtained in case of the software usage.

So now we stop here with this ANOVA table, and in the next lecture, we will try to consider the test of hypothesis on the individual regression coefficients, their confidence interval and some other aspect, till then goodbye.