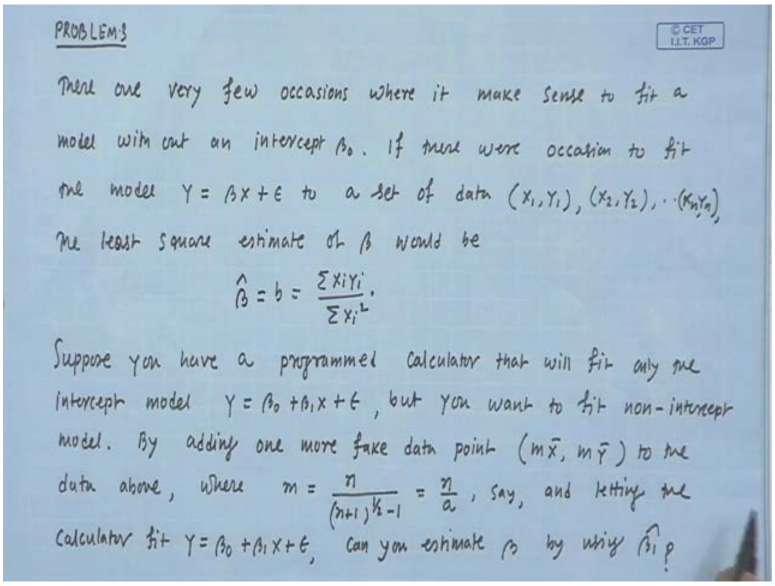
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, kharagpur

Lecture - 37
Tutorial – II

Hi, so this is my second tutorial class and we will be solving some problems from simple linear regression multiple linear regression. And also may be yield conditioning of the coefficient matrix x, so here is the problem.

(Refer Slide Time: 00:47)



PROBLEM:3

©CET I.I.T. KGP

There are very few occasions where it make sense to fit a model with out an intercept β_0. If there were occasion to fit the model $Y = \beta x + \epsilon$ to a set of data $(x_1, Y_1), (x_2, Y_2), \cdots (x_n, Y_n)$, the least square estimate of β would be

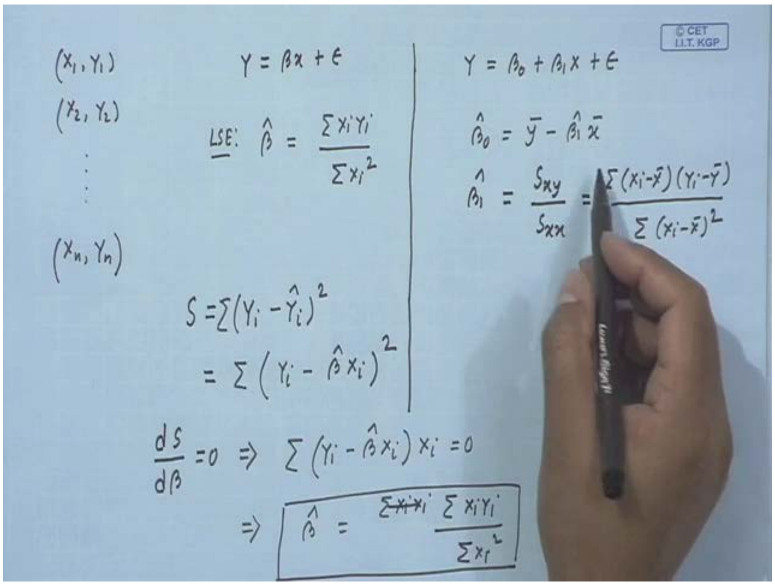
$$\hat{\beta} = b = \frac{\Sigma x_i Y_i}{\Sigma x_i^2}.$$

Suppose you have a programmed calculator that will fit only the intercept model $Y = \beta_0 + \beta_1 x + \epsilon$, but you want to fit non-intercept model. By adding one more fake data point $(m\bar{x}, m\bar{Y})$ to the data above, where $m = \frac{n}{(n+1)^{1/2} - 1} = \frac{n}{a}$, say, and letting the calculator fit $Y = \beta_0 + \beta_1 x + \epsilon$, can you estimate β by using $\hat{\beta}_1$?

Problem 3, because problem 1 and 2 we solved in the a previous tutorial, so this says that there are very few occasions where it makes sense to fit a model without an intercepted beta naught. If there were occasion to fit the model y equal to beta x plus epsilon that is a model without any intercept to a set of data x 1 y 1, x 2, y 2 x and y n and then least square estimate of this beta would be this. So, we will this part you know of what is the least square estimate for beta. Let me call it beta hat for a model without any intercept beta naught, and then the second part is that suppose you have a program calculator that will fit only the intercept model y equal to beta naught plus beta 1 x plus f silent.

But, you want to fit non intercept model, so you have some program calculator that fit only intercept model what you want to fit non intercept model. Now, the question is by adding one more fake data say m x bar m y bar, where m is a function of n and letting the calculator fit the intercept model can you estimate this beta by using beta 1 hat. So, this is the problem may be I will try to explain it once more the problem itself.

(Refer Slide Time: 02:59)



(x_1, y_1)

(x_2, y_2)

$\vdots$

(x_n, y_n)

$Y = \beta x + \epsilon$

LSE: $\hat{\beta} = \dfrac{\sum x_i Y_i}{\sum x_i^2}$

$S = \sum (Y_i - \hat{Y}_i)^2$

$= \sum (Y_i - \hat{\beta} x_i)^2$

$\dfrac{dS}{d\beta} = 0 \Rightarrow \sum (Y_i - \hat{\beta} x_i) x_i = 0$

$\Rightarrow \boxed{\hat{\beta} = \cancel{\sum x_i x_i} \dfrac{\sum x_i Y_i}{\sum x_i^2}}$

$Y = \beta_0 + \beta_1 x + \epsilon$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

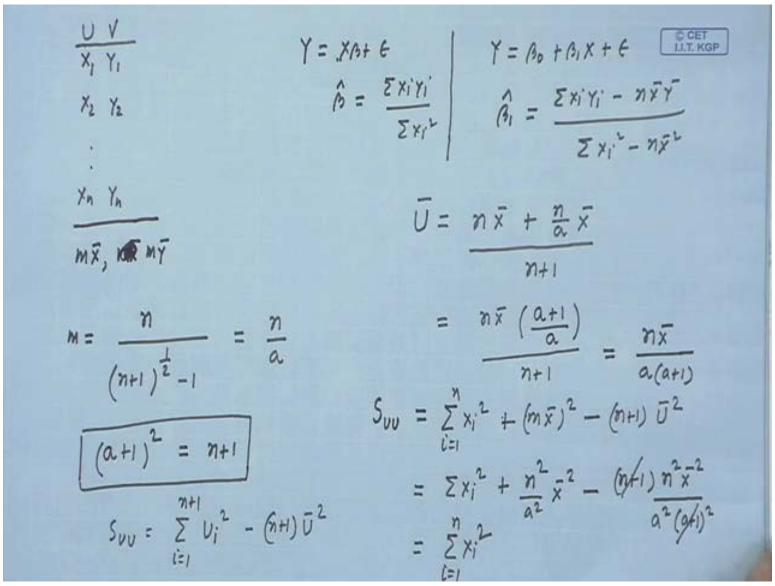
$\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

So, what we are given is that we are given a set of data x 2 y 1, x 2, y 2 and x n y n and you want to fit a non intercept model like y equal to beta x plus f silent. But, we have a program calculator for fitting this model with intercept that is y equal to beta naught plus beta 1 x plus f silent. So, we have a program to fit this model and we know that for this model beta naught is the estimate of beta naught is equal to y bar 1 hat x bar and beta 1 hat is equal to S x y by S x x. That means a this is x i minus x bar into y i minus y bar by summation x i minus x bar square, so we have a program to calculate these two things given a set of data.

But, what we want, we want to fit this model to the given data right and it says that the question says that the least square estimate for beta. Here, is equal to summation x i y i by summation x i square first check this one what is the least square estimate for beta in the non intercept model. So, to find the least square estimate for beta what we do is that we minimize this function s which is equal to y i minus y i hat this is the i-th residual and we know that this one is equal to y i minus y i hat i can replace by beta hat x i right. Then the least square estimate of beta hat is obtained by minimizing this function with respect to beta.

So, d S you differentiate it with respect to beta that equal to 0 implies that summation y i minus beta hat x i into x i equal to 0 which implies that beta hat is equal to summation x i y I, x i y i by x i square. So, we prove which you know the least square estimate for beta is a this quantity, now the problem is that we want to find this beta using the program calculators we have let me see the question once more. Suppose we have a program calculator that will fit only the intercept model this one, but we want to fit a non intercept model. The question is by adding one more fake data say this one, can we estimate beta by using beta one hat that means by using this program.

(Refer Slide Time: 07:58)



U V

X_1 Y_1

X_2 Y_2

$\vdots$

X_n Y_n

$m\bar{x}$, $m\bar{Y}$

$$m = \frac{n}{(n+1)^{\frac{1}{2}} - 1} = \frac{n}{a}$$

$$\boxed{(a+1)^2 = n+1}$$

$$S_{UU} = \sum_{i=1}^{n+1} U_i^2 - (n+1)\bar{U}^2$$

$$Y = X\beta + \epsilon$$

$$\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$$

$$\bar{U} = \frac{n\bar{x} + \frac{n}{a}\bar{x}}{n+1} = \frac{n\bar{x}\left(\frac{a+1}{a}\right)}{n+1} = \frac{n\bar{x}}{a(a+1)}$$

$$S_{UU} = \sum_{i=1}^{n} x_i^2 + (m\bar{x})^2 - (n+1)\bar{U}^2$$

$$= \Sigma x_i^2 + \frac{n^2}{a^2}\bar{x}^2 - \cancel{(n+1)}\frac{n^2\bar{x}^2}{a^2\cancel{(a+1)^2}}$$

$$= \sum_{i=1}^{n} x_i^2$$

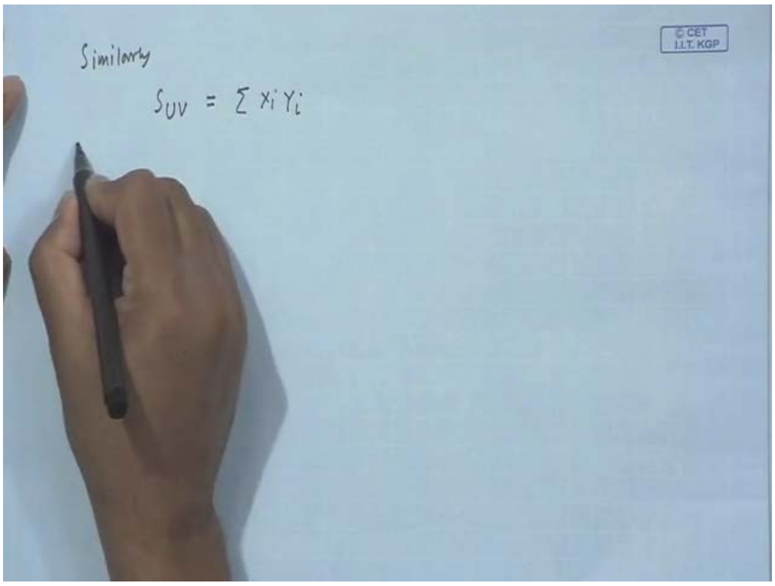
So, at this moment we have the data x i, x 1 y 1, x 2 y 2 and x n y n and if you add one more data say n plus 1 of that is m x bar n, sorry m x bar and m y bar where m is equal to m is equal to m by n plus 1 to the power of half minus 1. This is equal to say n by a then what is a, a is from, here I can write that a plus 1 square is equal to n plus 1, now what we will do is that we want to estimate x beta plus f silent, we want to estimate beta hat which is equal to x i y i by x i bar. But, we do not have program, for this one we have program for estimating the model intercept model that is y equal to beta naught plus beta 1 x f silent.

The program calculator gives us beta 1 hat equal to summation x i y i minus n x bar y bar by summation x i square n x bar square. Now, we used this formula for the revised data let me call this data as u v, now for this new data which is, now in involving n plus 1 points what is u bar of u bar is equal to n x bar plus a new data m x bar m is nothing. But, n by a x bar by n plus 1 this one is equal to of n x bar into a plus 1 by a by n plus 1.

So, we know that n plus 1 is a plus 1 whole square, so I can write it this way, so I can write this one as n x bar by a into a plus 1. So, this is my u bar for the data involving n plus 1 points and now let me compute S u u, S u u is nothing but you know what is S x x. So, S u u is equal to of summation x i square plus the new data m x bar square minus n plus 1 u bar square because s u u is a let me write. Here, s s u is nothing but summation u i square minus n plus 1 u bar square and here i is from 1 to n plus 1, so that is what I wrote, here and here i is from 1 to n and then we see the n plus 1 data.

This 1 can written as summation x i square plus n square by a square into x square minus n plus 1 u bar square what is u bar square, u bar is this quantity. So, n square x bar square by a square into a plus 1 square, so I can cancel out these two things and then you can see that this term same as this term. So, you are left with summation x i square, so S u u is equal to x i square from i equal to 1 to n.

(Refer Slide Time: 14:35)

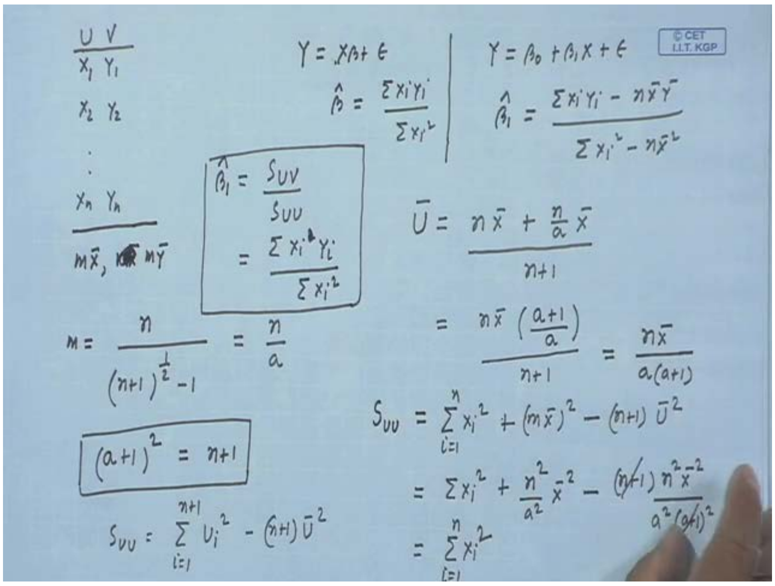


Similarly

$$S_{UV} = \sum x_i Y_i$$

Similarly, you can prove that, similarly you can prove that S u v is equal to summation x i y i.

(Refer Slide Time: 15:05)



U	V
X_1	Y_1
X_2	Y_2
$\vdots$	
X_n	Y_n
$m\bar{X}$,	$m\bar{Y}$

$$Y = X\beta + \epsilon \qquad \hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$$

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \hat{\beta}_1 = \frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum x_i^2 - n\bar{x}^2}$$

$$\boxed{\hat{\beta}_1 = \frac{S_{UV}}{S_{UU}} = \frac{\sum x_i Y_i}{\sum x_i^2}}$$

$$m = \frac{n}{(n+1)^{\frac{1}{2}} - 1} = \frac{n}{a}$$

$$\boxed{(a+1)^2 = n+1}$$

$$S_{UU} = \sum_{i=1}^{n+1} U_i^2 - (n+1)\bar{U}^2$$

$$\bar{U} = \frac{n\bar{x} + \frac{n}{a}\bar{x}}{n+1} = \frac{n\bar{x}\left(\frac{a+1}{a}\right)}{n+1} = \frac{n\bar{x}}{a(a+1)}$$

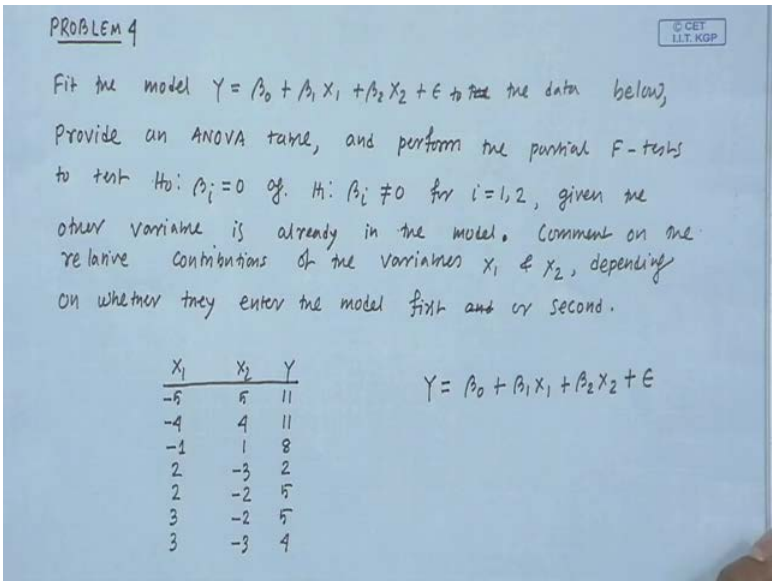
$$S_{UU} = \sum_{i=1}^{n} x_i^2 + (m\bar{x})^2 - (n+1)\bar{U}^2$$

$$= \sum x_i^2 + \frac{n^2}{a^2}\bar{x}^2 - \cancel{(n+1)}\frac{n^2\bar{x}^2}{a^2\cancel{(a+1)^2}}$$

$$= \sum_{i=1}^{n} x_i^2$$

So, now if you apply the program calculator to the new data new set of data which involves n plus 1 data points and we called them u v then the program calculator will give you the estimate beta 1 hat which is equal to S u v by S u u. You have proved that S u v is nothing but summation x i square, sorry summation x i y i and S u u is equal to summation x i square. So, you prove that you know using the program calculator if you use a, if you add a fake data, here at the end then you can estimate the model without intercept using a formula of model including the intercept. So, that is what we proved just now.

(Refer Slide Time: 16:24)



PROBLEM 4

Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to ~~fit~~ the data below, Provide an ANOVA table, and perform the partial F-tests to test $H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$ for $i = 1, 2$, given the other variable is already in the model. Comment on the relative contributions of the variables X_1 & X_2, depending on whether they enter the model first ~~and~~ or second.

X_1	X_2	Y
−5	5	11
−4	4	11
−1	1	8
2	−3	2
2	−2	5
3	−2	5
3	−3	4

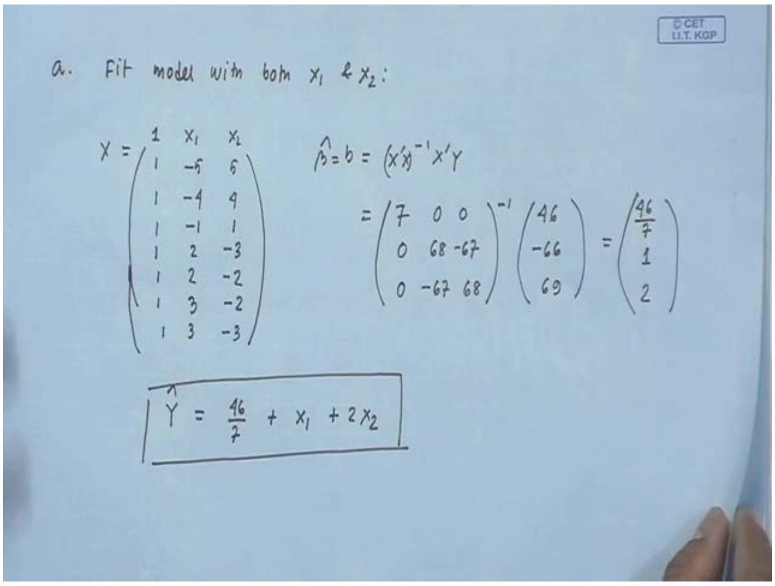
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Now, let me consider another problem, so called this problem 4, so this one says that fit the model y equal to beta naught plus beta 1 x 1 plus beta 2 x 2 plus epsilon to the data given below. So, here we have two regressors x 1 and x 2 and one response variable y and what it to do is that you provide and ANOVA table. This is quite set forward problem and perform the partial f tests to test h naught that beta i equal to 0 against h 1 that beta is not equal to 0 for i equal to 1 2 given the other variable is already in the model. So, test the significance of i-th regressors in the presence of other regressors in the model.

So, let me do till this one and then it set the comment on the relative contribution of the variables x 1 and x 2 depending on whether they enter the model first or second I will come to this point later on. So, it is a multiple linear regression problem involving two regressors and you are given a data like for x 1 x 2 and y. So, you have to fit the model that means you have to estimate the parameters beta naught beta 1 and beta 2.

So, once you done with the estimation of the parameters then we can have a ANOVA table and after getting the ANOVA table we estimate, sorry we test the significance of the model. That is called the global test and after the global test what we will do we will go for the partial f tests to test the significance of the i-th regressors in the presence of other regressors. After that we just short of compare the relative contribution of x 1 and x 2, I mean which is more significance to explain the variability in y. So, this is the given data for two regressors and one response variable, and we have to fit the model, the model is y equal to beta naught plus beta 1 x 1 plus beta 2 x 2 plus epsilon and we know how to fit this model.

(Refer Slide Time: 19:46)



© CET I.I.T. KGP

a. Fit model with both x_1 & x_2:

$$X = \begin{array}{c} \begin{array}{ccc} 1 & x_1 & x_2 \end{array} \\ \begin{pmatrix} 1 & -5 & 6 \\ 1 & -4 & 4 \\ 1 & -1 & 1 \\ 1 & 2 & -3 \\ 1 & 2 & -2 \\ 1 & 3 & -2 \\ 1 & 3 & -3 \end{pmatrix} \end{array}$$

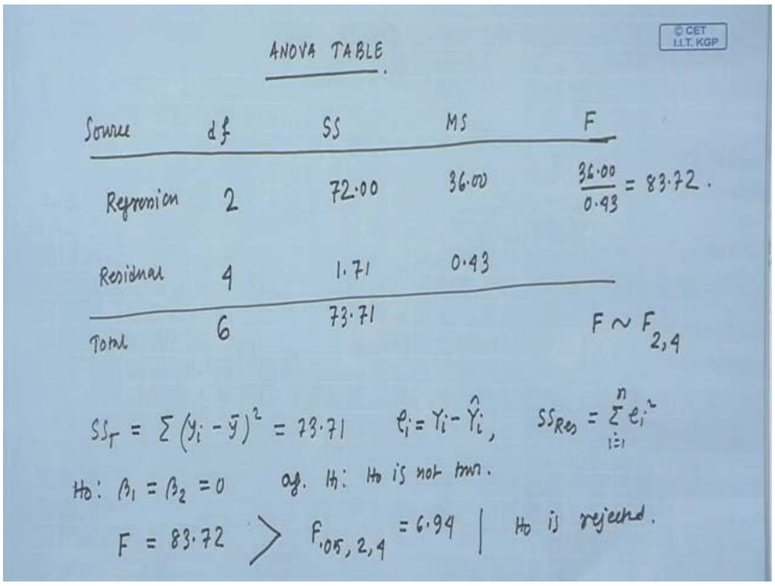
$$\hat{\beta} = b = (X'X)^{-1}X'Y$$

$$= \begin{pmatrix} 7 & 0 & 0 \\ 0 & 68 & -67 \\ 0 & -67 & 68 \end{pmatrix}^{-1} \begin{pmatrix} 46 \\ -66 \\ 69 \end{pmatrix} = \begin{pmatrix} \frac{46}{7} \\ 1 \\ 2 \end{pmatrix}$$

$$\boxed{\hat{Y} = \frac{46}{7} + x_1 + 2x_2}$$

So, what I will do is that, so first we will write down the x matrix, so x matrix you can see that this is corresponds to x naught which all 1 and the x 1 and x 2. So, once you have a x matrix you know that the estimated beta hat is equal to x prime x inverse x prime y. So, you are given y, you know x, you can compute you can check that beta naught is equal to 46 by 7, beta 1 is equal to 1 and beta 2 is equal to 2. So, the fitted model is this one, this is the fitted model y hat is equal to 46 by 7 plus x 2 beta 1, x y beta 1, x 1 beta 1 is equal to 1 and beta 2 x 2 beta 2 is equal to 2, so this is the fitted model.

(Refer Slide Time: 21:14)



© CET I.I.T. KGP

ANOVA TABLE.

Source	df	SS	MS	F
Regression	2	72.00	36.00	$\frac{36.00}{0.43} = 83.72.$
Residual	4	1.71	0.43	
Total	6	73.71		$F \sim F_{2,4}$

$SS_T = \sum(y_i - \bar{y})^2 = 73.71$ $\quad e_i = Y_i - \hat{Y}_i,$ $\quad SS_{Res} = \sum_{i=1}^{n} e_i^2$

$H_0: \beta_1 = \beta_2 = 0$ $\quad$ ag. H_1: H_0 is not true.

$F = 83.72 > F_{.05, 2, 4} = 6.94$ | H_0 is rejected.

Now, what we have to do is that once we have the a fitted model we will go for the a ANOVA table, so ANOVA table, so the source and then degree of freedom, sum of square M S and F value and the sources are, here the total. So, what is S S total S S T is a equal to summation y i minus y bar square and you can check that you are you know the y i values, so you can check that this is equal to 73.71. So, my S S total is 73.71 and we know y hat i values and then we know the original observed values also y i. So, from here you can compute e i, the i-th residual and then you can compute the S S residual also S S residual is nothing but summation e i square i is from 1 to n, so you can check that the S S residual is 1.71.

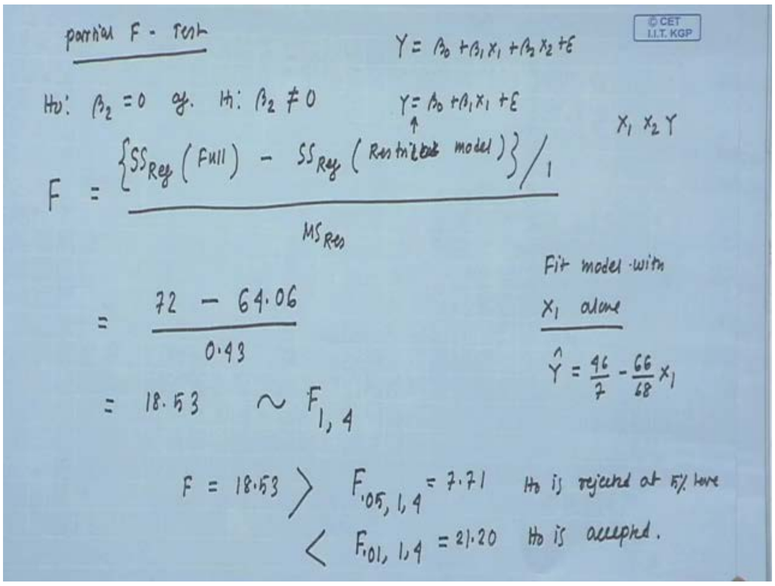
So, this is residual and then we left with the regression, so the regression you can check that the regression is 72.00, now here is the problem how many observations we have, we have a 7 observations. So, the degree of freedom for S S total is 6 and then the residual degree of freedom would be 4 because there are 7 e i s and there are 3 restriction because of 3 parameters. So, the degree of freedom for the residual is 4 and the regression has the degree of freedom 2. So, now you can compute the M S value, M S values are 336, here and this is 0.43, so the F statistic value is 36.00 by 0.43 which is equal to 83.72, so using this F value this, is this is the total variability of in the response about the mean.

This is a part of F total variability which is explained by this model and this is a part which remains unexplained. So, we can test the significance of the fitted model by testing this hypothesis that h naught that beta 1 is equal to beta 2 is equal to 0. So, this says that a fitted model is not significance, not significant against the alternative hypothesis which is say that H naught is not true that means the alternative hypothesis is what is this says that the fitted model is significance. You can clearly see that you know the model is very significant because it almost 99 percent of that total variability explained by the fitted equation.

So, how to test this one, this one can be test using the F statistic given here, so the observed F value which is equal to 83.72 and you compare with, compare this value with the tabulated value what is the degree of freedom for F. So, F follows, here F has degree of freedom 2, 4, now you check the tabulated value of F 0.05, 2, 4 from the F table that you can check that this one is 6.94, so the observed value is greater than the tabulated value which implies that H naught is rejected.

So, the global test says that the fitted model is significant, now what we will do is that there are two regressors variable in the model we will test the significance of say x 2 fast in the presence of x 1. So, we will test whether x 2 is significant in the presence of x 1 when x 1 is there in the model and then again, similarly what we will do is that we will test the significance of x 1 in the presence of x 2 in the model. So, those things we will do using of partial F test also we can go for a t test.

(Refer Slide Time: 28:16)



partial F - Test

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

$H_0: \beta_2 = 0$ ag. $H_1: \beta_2 \neq 0$

$Y = \beta_0 + \beta_1 x_1 + \epsilon$

$x_1 \; x_2 \; Y$

$$F = \frac{\{SS_{Reg}(\text{Full}) - SS_{Reg}(\text{Restricted model})\}/1}{MS_{Res}}$$

$$= \frac{72 - 64.06}{0.43}$$

$$= 18.53 \sim F_{1,4}$$

Fit model with x_1 alone

$\hat{Y} = \frac{46}{7} - \frac{66}{68} x_1$

$F = 18.53 > F_{.05,1,4} = 7.71$ H_0 is rejected at 5% level

$< F_{.01,1,4} = 21.20$ H_0 is accepted.

Now, we will go for partial F test, so we test H naught that beta 2 is equal to 0 against the H 1 that beta 2 is not equal to 0 and this test is in the presence, in the presence of x 1 in the model, so how do you test this one we go for if you go for the partial F test. Here, is the F statistic, F is equal to S S regression for the full model minus S S regression for the restricted model, restricted model. We will see you can check that this has to be divided by 1 by M S residual we know what is the full model, full model is y equal to beta naught plus beta 1 x 1 plus beta 2 x 2 plus epsilon and a restricted model is the model under H naught.

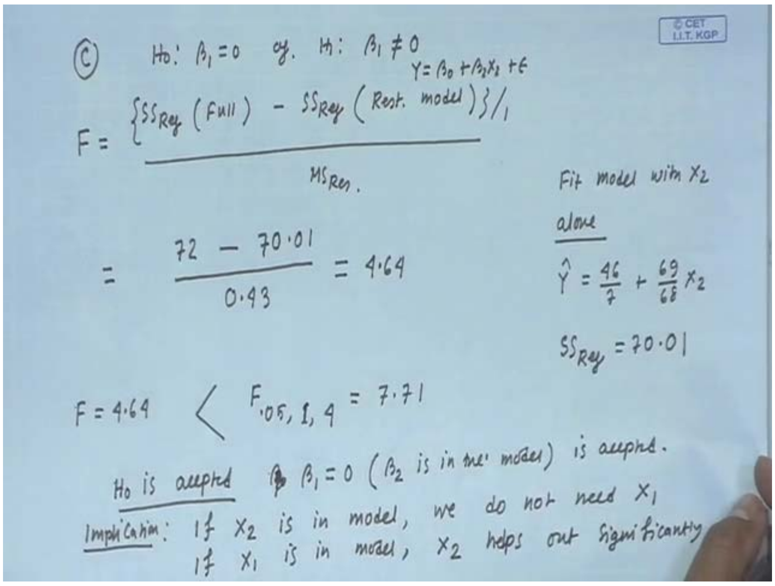
So, restricted model is basically y equal to beta naught plus beta 1 x 1 plus F silent, so we know what is S S regression for the full model. So, if you look at the ANOVA table we know that S S regression for the full model is 72, so here we will put 72 minus. Now, to find the S S regression for this restricted model what you have to do is that you have to fit a model with x 1 alone that means you have to fit this model.

You can check that the fitted model would be y hat is equal to 46 by 7 minus 66 by 68 into x 1, so you have you are given the data x 1 x 2 y, so you can fit a model between x 1 and y you know how that that means a simple linear regression. So, this is a fitted model and once you have a fitted model you can compute the S S regression due to this model that that is nothing but 64.06 you check this one. Now, this has degree of freedom 2 and this has degree of freedom 1 I hope you know why that is, so 2 minus 1 is equal to 1, so you divide this by one and now the M S residual from the ANOVA table of M S residual is 0.43.

So, you put that the M S residual is 0.43 well, so this one is equal to 18.53 and we know that this F statistic has degree of freedom 1, 4, 4 is the residual freedom. Now, you find the tabulated value of F 0.05, 1, 4 that is equal to 7.71 and you have observed value F that is 18.53. So, this test says that yes beta 2 is significant, so this means H naught is rejected, so H naught is rejected at 5 percent level of significance and of, so at the 5 percent level of significance we can say that x 2 is significance in the presence of this one in the model.

Now, let us check whether this is significant at a 0.01 level of significance, so compute you find the value of tabulated value of F 0.01, 1, 4 that you can check that this is equal to 21.02. So, the F value is less than this one, so here H naught is accepted that means at 1 percent level of significance x 2 is not significant in the presence of x 1. Whereas, at the 5 percent level of significance x 2 is significant in the presence of x 1, so that is the conclusion of this partial test. So, at least you know at 5 percent significance we observed that beta 2 is significant or x 2 is significant in the presence of x 1.

(Refer Slide Time: 34:50)



© CET I.I.T. KGP

ⓒ $H_0: \beta_1 = 0$ agt. $H_1: \beta_1 \neq 0$

$Y = \beta_0 + \beta_2 x_2 + \epsilon$

$$F = \frac{\{SS_{Reg}(\text{Full}) - SS_{Reg}(\text{Rest. model})\}/1}{MS_{Res}.}$$

$$= \frac{72 - 70.01}{0.43} = 4.64$$

Fit model with x_2 alone

$$\hat{Y} = \frac{46}{7} + \frac{69}{68} x_2$$

$SS_{Reg} = 70.01$

$F = 4.64 < F_{.05, 1, 4} = 7.71$

H_0 is accepted $\beta_1 = 0$ (β_2 is in the model) is accepted.

Implication: If x_2 is in model, we do not need x_1

If x_1 is in model, x_2 helps out significantly

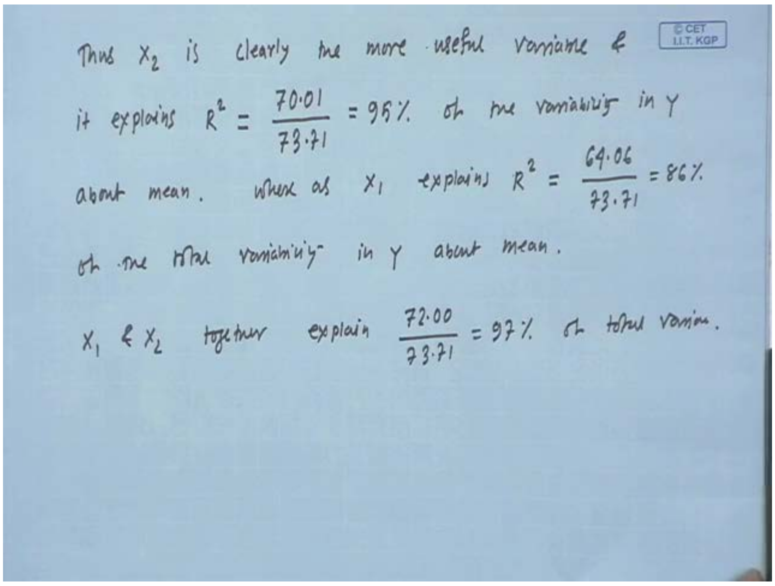
Now, what we will do is that we will check the significance of x 1 in the presence of x 2, so we will go for the partial test say H naught that is beta 1 equal to 0 against the alternative hypothesis H 1 that beta 1 is not equal to 0. So, here same thing we the statistic for testing this hypothesis F, which is S S regression for the full model minus S S regression for the restricted model. Here, it should be divided by 1 by M S residual, so here what is the restricted model.

Here, the restricted model is y equal to beta naught plus beta 2 x 2 plus epsilon because the restricted model is the model under H naught. So, again you have to fit model with x 2 alone and you can check that fitted model is y hat equal to 46 by 7 plus 69 by 68 x 2 and once you have the fitted model you can find out the S S regression. So, S S regression is equal to 70.01, so this one is greater than the S S regression we got for the model with x 1 alone that was 64 anyway.

So, this one is equal to the 72 is a regression, S S regression for the full model and this one is 70.01 by M S residual we know that is 0.43. So, this one is equal to 4.64, now you check, so your observed value is 4.64 and the tabulated value F 0.05, 1, 4, here that is equal to 7.71. The observed value is less than the tabulated value that means H naught is accepted, what is the meaning of this one that means beta naught 1, sorry that means beta 1 equal to 0 when beta 2 is in the model beta 1 equal to 0 is accepted. So, the implication of this one is that if x 2 is there in the model we do not need x 1 because you can see that the full model can explain 72, I mean for the full model S S regression is 72.

For the model involving only x 2 is 70, so which is almost like full model, so the implication of this one is that implication is that if x 2 is in model we do not need x 1 x 2 is enough. So, if you have x 2 first in the model we do not need to include x 1, but if x 1 is in model then we have tested that you know the significance of beta 2 in the presence of x 1 is significant. So, if this one is in the model then we include x 2, so x 2 helps out significantly, so this is what the implication of these two partial test.

(Refer Slide Time: 41:08)



Thus X_2 is clearly the more useful variable & it explains $R^2 = \frac{70.01}{73.71} = 95\%$ of the variability in Y about mean. Where as X_1 explains $R^2 = \frac{64.06}{73.71} = 86\%$ of the total variability in Y about mean.

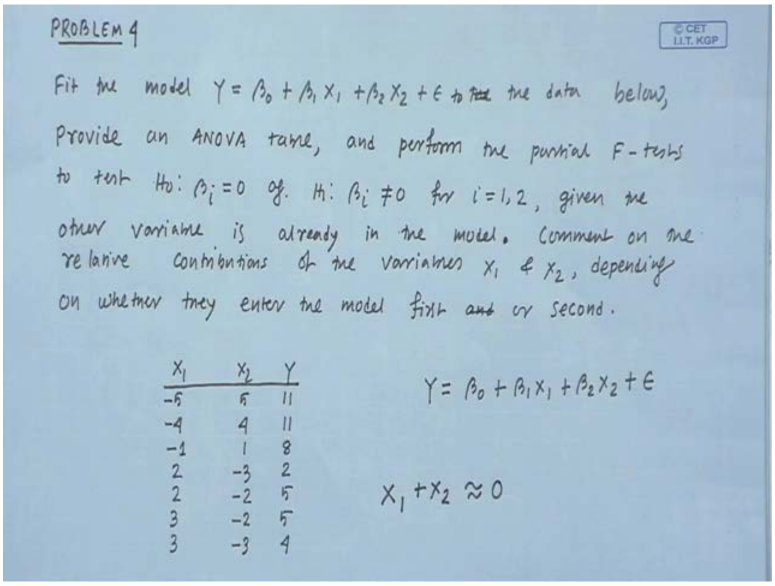
X_1 & X_2 together explain $\frac{72.00}{73.71} = 97\%$ of total varian.

© CET I.I.T. KGP

Let me also conclude that thus x 2 is clearly the more useful variable and it explains I will compute the coefficient of determination R square for x 2 itself. So, for the model involving x 2 alone we have observed that S S regression is 70.01 and the total variability is in the response variable is put S S T is 73.71. That means the 95 percent of the total variability is explains by x 2 alone, so 95 percent of the variability in y it explains 95 percent of the variability in y about mean where as x 1 explains let me compute R square.

So, along with s only x 1 in the model the S S regression was 64.06 by S S T that is 73.01 this is 86 percent, so x 1 explain 86 percent of the total variability in y about mean. So, x 1 explain 95 percent of the total variability in y, whereas sorry, x 1 explain 95 percent of the total variability in y, whereas x 1 explains 86 percent of the total variability in y and x 1 and x 2 together explain 72 by 73.71 that is 97 percent of total variability. So, the conclusion is that x 2 is a more useful regressor variable than x 1 because x 2 alone can explain 95 percent of total variability and also x 2 is significant in the presence of x 1. Whereas, x 1 explain 86 percent of the total variability and x 1 is not significant in the presence of x 2.

(Refer Slide Time: 45:07)



PROBLEM 4

© CET I.I.T. KGP

Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to ~~fit~~ the data below, Provide an ANOVA table, and perform the partial F-tests to test $H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$ for $i = 1, 2$, given the other variable is already in the model. Comment on the relative contributions of the variables X_1 & X_2, depending on whether they enter the model first ~~and~~ or second.

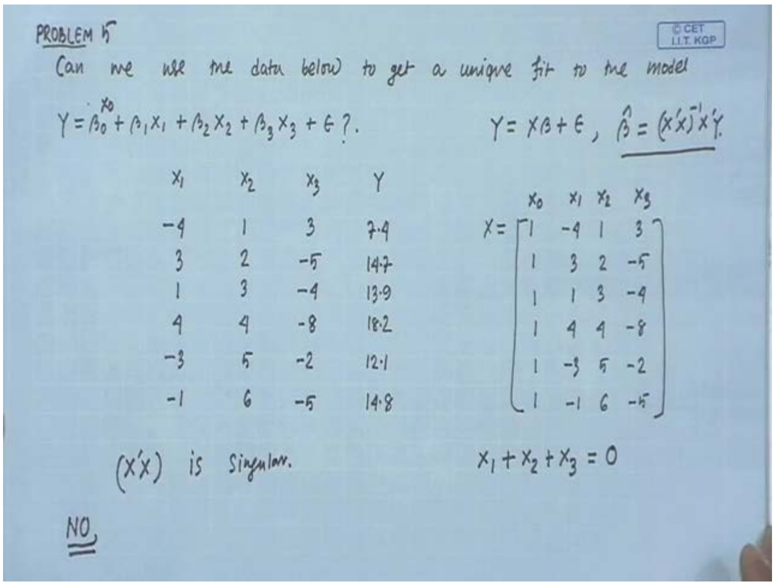
X_1	X_2	Y
-5	5	11
-4	4	11
-1	1	8
2	-3	2
2	-2	5
3	-2	5
3	-3	4

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$X_1 + X_2 \approx 0$$

But, one more thing you know I just want to say this problem is particularly interested interesting because of the one more fact. Here, you can see that the data, here you can see data there x 1 and x 2 are not independent I mean x 1 can be written in terms of x 2, in fact you can check that x 1 plus x 2 is almost equal to 0. So, there two regressor variable are dependent, so this short of indicates that you know there could be multi collinear I mean in fact there is multi collinear, co linearity in the this data. That is why the test also says that you know you do not need x 1 if x 2 is present in the model. So, it is enough to keep only x 2 in the model I mean both x 1 and x 2 are not required, so next will be considering another problem.

(Refer Slide Time: 46:39)



PROBLEM 5

© CET I.I.T. KGP

Can we use the data below to get a unique fit to the model

$Y = \beta_0 \overset{x_0}{} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$?.

$Y = X\beta + \epsilon$, $\underline{\hat{\beta} = (X'X)^{-1} X'Y}$.

X_1	X_2	X_3	Y
-4	1	3	7.4
3	2	-5	14.7
1	3	-4	13.9
4	4	-8	18.2
-3	5	-2	12.1
-1	6	-5	14.8

$$X = \begin{array}{c} \begin{array}{cccc} x_0 & x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1 & -4 & 1 & 3 \\ 1 & 3 & 2 & -5 \\ 1 & 1 & 3 & -4 \\ 1 & 4 & 4 & -8 \\ 1 & -3 & 5 & -2 \\ 1 & -1 & 6 & -5 \end{bmatrix} \end{array}$$

$(X'X)$ is singular.

$X_1 + X_2 + X_3 = 0$

NO,

So, this problem is a let it call me problem 5, so it says that can we use the data below to get a unique fit to the model y equal to beta naught plus beta 1 x 1 plus beta 2 x 2 plus beta 3 x 3 plus f silent. So, can we fit this model uniquely using this data that is the question, so look at the data carefully, so this involves how many parameters it has a 1, 2, 3, 4 parameter and we have this data. So, it is a multiple linear regression model with three regressors, so you can write it in this form y equal to x beta plus epsilon and then we know that, we know how to estimate this regression coefficient uniquely.

We know that beta hat is equal to x prime x inverse x prime y and why what is the problem, here then why it is says that can we use the data below to get a unique fit to the model why not. So, look at the x matrix, here I have included x naught, here x naught means beta naught x naught and then x 1, x 2 and x 3 and you know that x 1 column is correspond to this column. Then simply I will compute x prime x inverse and then I get the estimate, but is there any problem here, so we need to check whether they exist or no all that.

Whether this all this columns are independent or not, here I can see that x 1 plus x 2 plus x 3 is equal to 0 that means a the columns of this matrix are not independent which implies that x prime x is singular and then if it is. So, the determinant of x prime x is going to I mean its singular, so determinant of x prime x is going to be 0, so you cannot compute the inverse of this x prime x matrix. So, that is why the problem you know, you cannot compute the beta uniquely here, so the ultimate answer to this question is no. So, we cannot use the data below to get a unique fit to this model, so this problem is related to the yield conditioning of x matrix.

So, today we considered 3 problems you know first problem was from the simple linear regression model, the second problem was base standard problem in multiple linear regression model. The problem was very interesting you know it is a short of you have two regressors and then we, finally we observed that you know x 2 is significant in the presence of x 1. So, but whereas x 1 is not significant in the presence of x 2, so x 2 is more useful regressor variable than x 1 for the given data and finally we observed the two regressors are not independent, so there exist multi collinearity. So, that is why one variable is enough to explain the total variability in the response variable and then finally the fifth problem was about the yield conditioning of the coefficient matrix. So, in the next tutorial again you know we will discuss some more problem.

Thank you.