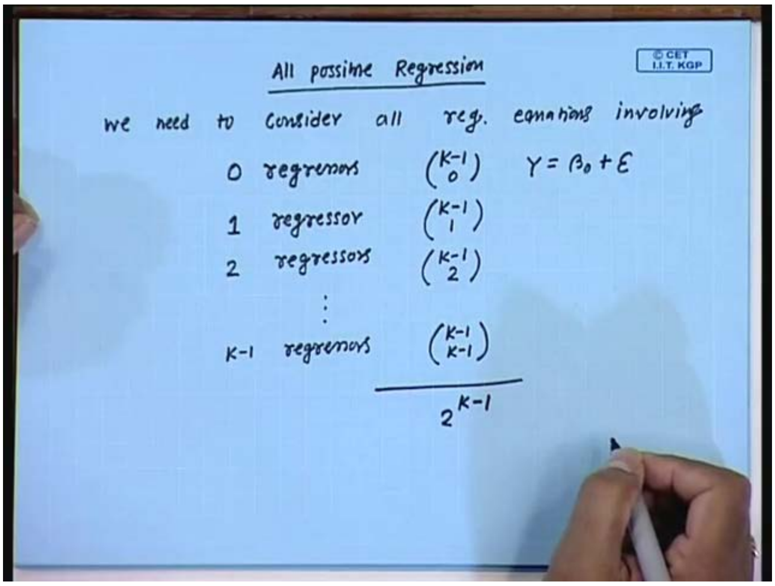
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 10
Selecting the Best Regression Model

Hi, this is my first lecture on Selecting the Best Regression Model, we are under the multiple linear regression set of, and you know that you know in multiple linear regression, the number of regression variables is more than one. And in most of the practical problems what happened is that, you know the number of regression is very large and having the large number of regression variables. We may wonder you know whether a some of them can be are irrelevant and can be removed from the regression equation.

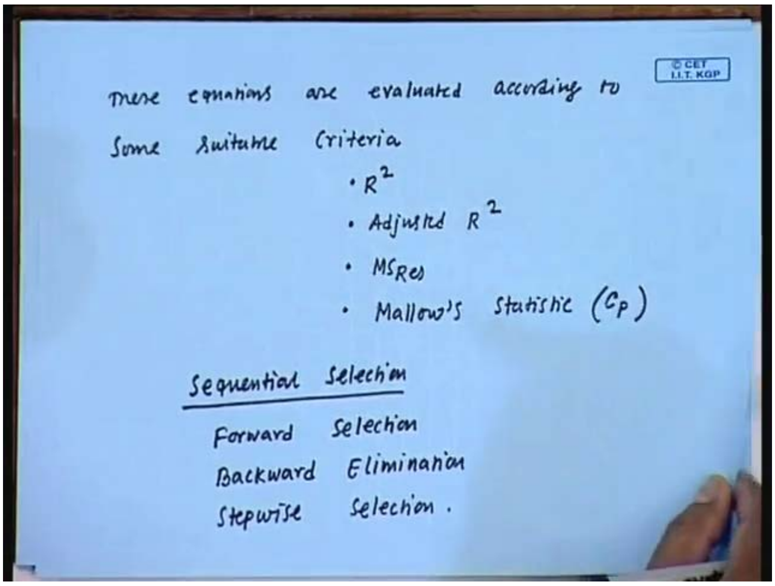
So, the basic idea you know behind this finding the best regression model is that, we need to find an appropriate substrate of regressors that can explain the variability in the responsible variable well. And finding this substrate regression variable, this problem is called variable selection problem well, let me explained the thing in detail, there are several algorithm to solve this problem. And those algorithm can be you know divided to I mean that can be classified into two classes basically, one approach is called all possible regression approach, and the other one is called sequential selection well. So, first I will be talking about all possible regression.

(Refer Slide Time: 03:16)



All possible regression, say here you know we need to consider all regression equations involving say 0 regresses well. So, if is there are k minus 1 is the total number of regresses in multiple linear regression model, then you know number of model have been 0 regresses k minus 1 see 0 in the model is basically Y equal to beta naught plus epsilon. So, we will also consider I mean of course, the regression equations are models involving 1 regressor and the number of models number of such models is k minus 1 see 1 to 2 regressors k minus 1 to see 2 the regression model or there involving to regressor variables well. Similarly, we go up to k minus 1 regresses, so number of models involving k minus 1 regresses is 1, so total we have 2 to the power of k minus 1 regression models.

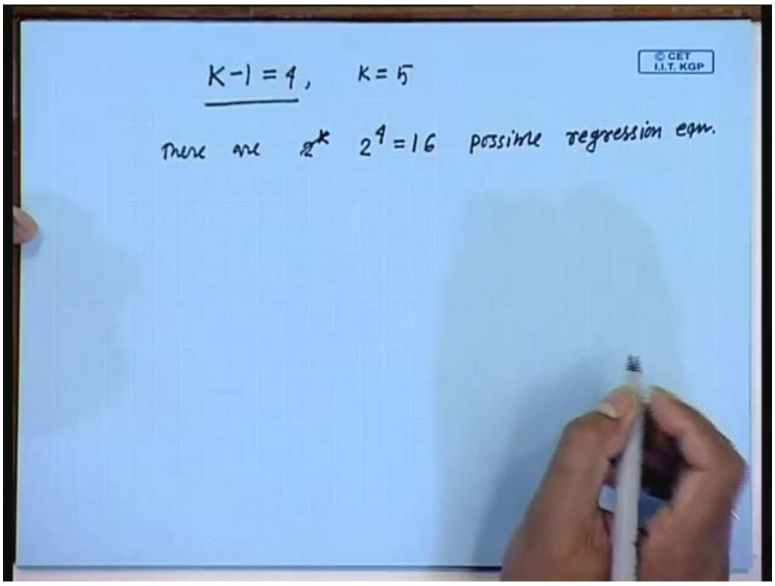
(Refer Slide Time: 06:06)



And you know this models these equations are evaluated according to some suitable criteria, first one is called R square this is the coefficient of multiple determination or coefficient of determination. And then we will be talking about the criteria adjusted R square, and then MS residual in the finally, I mean will evaluate the equations based on the criteria, mallows statistic. And this one is denoted by CP, and this is you know one approach that is you know all possible regression, and the other approach is called sequential selection.

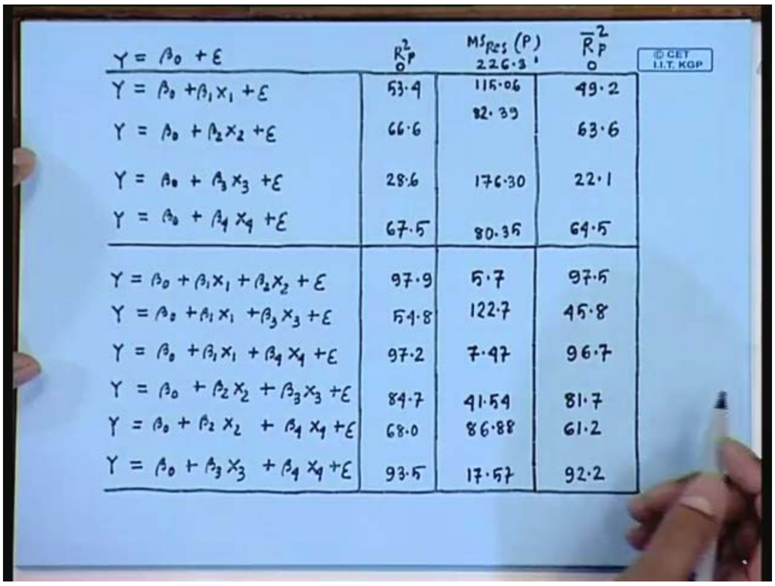
So, I will be talking about this sequential selection later on, and there are three algorithms of this type, those are called forward selection, back ward elimination, and the stepwise selection. So, today we will be talking about you know this all possible regression and how to evaluate, so many I mean 2 to the power of k minus 1 regression equation, based on this criteria well.

(Refer Slide Time: 09:05)



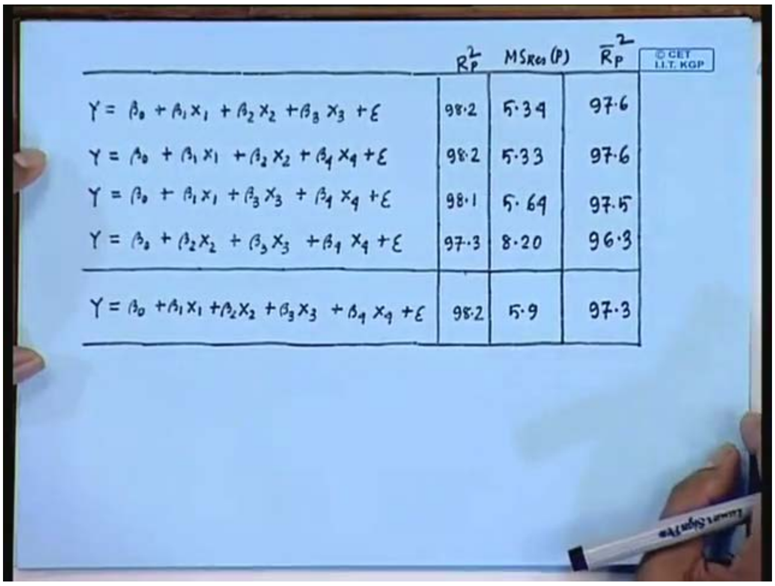
Now, if a the number of regresses is 4, so usually denote the number of regressors by k minus 1. So, if K minus 1 is equal to 4 K basically you know the K denotes the number of unknown parameters in the model well, so if there are K minus 1 regressors there will be K minus 1 regression coefficient. And this another unknown parameters the intercept, so the total we will have K unknown parameters well. So, if there are 4 regressors the problem then there are 2 to the power of 4 which is equal to 16 possible regression equations.

(Refer Slide Time: 10:38)



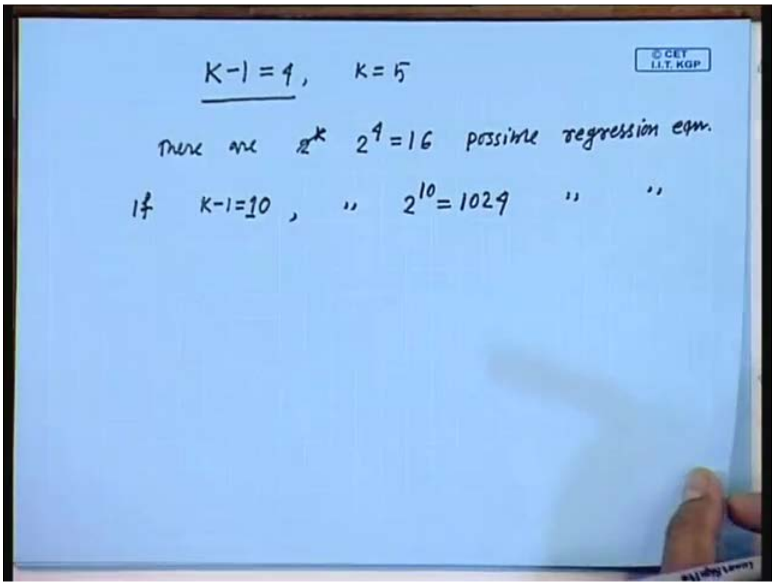
And let me just I have those 16 you know regression equation, so here I am concentrating a the problem with 4 regression variable. So, this is the model which without an no regression variable, so number of such model is 4 see 0 which is equal to 1, now these are the model involving 1 regression variable. So, this 1 is involving x 1, the second question is involving x 2, x 3 and x 4, so this are four the regression models involving 1 regression variable. And then we have you know six regression model involving two regression variable. So, this one is involving x 1, x 2, x 1, x 3, x 1, x 4, x 2, x 3, x 2, x 4, x 3, x 4.

(Refer Slide Time: 12:07)



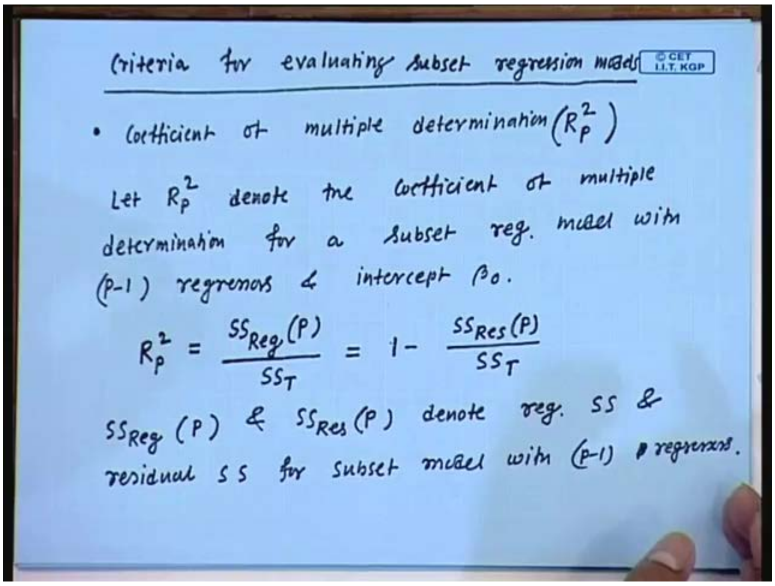
And then next we have regression model, involving three regression variable, So there are 4 see three that is equal to 4 such models regression model. So, this one involving x 1, x 2, x 3 like that and this is basically full model, and this involves all the 4 regression variable. So, there are 4 see 4; that means, 1 such model.

(Refer Slide Time: 12:45)



So, when the number of regression variable is 4 we have know 16 regression models, and we need to evaluate them with respect to some criteria. And see the complexity of this approach if you have problem with say K minus 1 equal to 10; that means, the number of regressor is equal to 10, then there are you know 2 to the power of 10 which is equal to 1024 possibility regression equation. So, clearly you know the number of questions are the number of models that need to be fitted you know that increases rapidly with the number of regression well. So, but still you know since in most of the practical problems, the number of regression variable could be like 22, 30, so but of course, you can use computed to fit all possible 2 to the power of 20 models also there is no problem well.

(Refer Slide Time: 14:26)

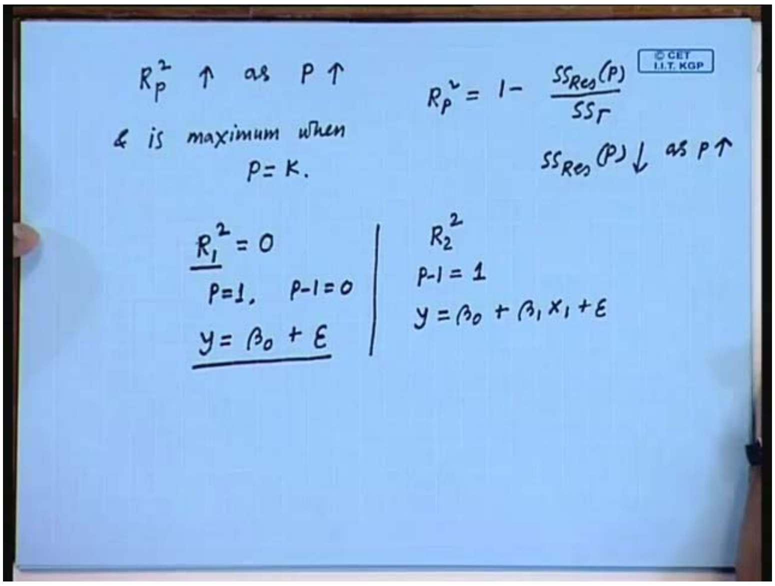


So, next I will be talking about the criteria, the first criteria it is mention that criteria for evaluating subset regression model well. So, we need to evaluate those subset models and the first criteria to evaluate them is I mean one criteria is coefficient of multiple determination, and we denote this one by R square. So, before also I told I mention about R square, and we used to call it like coefficient of determination, and hence since we talking about you know multiple linear regression model here, we call it multiple coefficient of multiple determination.

So, we denote this by RP square is let RP square denote the coefficient of multiple determination for a subset regression model with P minus 1 regresses and intercept beta naught well. So, by RP square you know this P is basically stands for the number of unknown parameters in the model, so since there are P minus 1 regressors there will be P minus 1 coefficient. And the intercept beta naught, the total number of unknown parameters is equal to P, and we denote the corresponding coefficient of the multiple determination by RP square.

So, this RP square is equal to SS regression P by SS T, which can be written as 1 minus SS residual P by SS T right. So, what is this SS regression P and SS residual P the denote regression SS, and residual SS for subset model with P minus 1 regressor, and so basically the RP square is associated with the model, when there are P minus 1 the regresses in the model. And RP square is parameter, which measures the proportion of variability in the response variable, which is explained by the regression model involving P minus 1 regresses well.

(Refer Slide Time: 20:11)



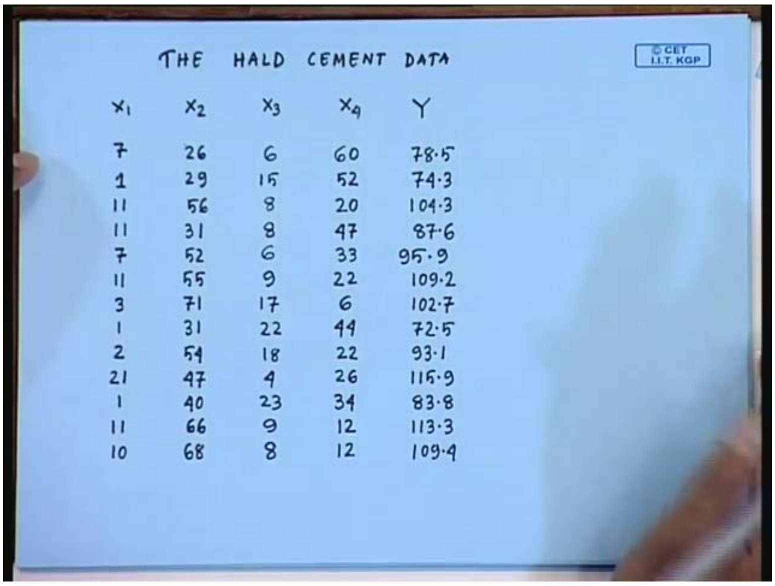
So, it is not I mean like you know this RP square it increases as P increases because you look at the definition of RP square, RP square is equal to 1 minus SS residual P by SS T. And we know that, SS residual this decreases, these decreases as P increases, so from here you know we can easily observe that RP square increases as has been increases. And this is maximum, when P call to K because you know P call to K means, P minus 1 is equal to K minus 1.

That means, we talking about the full model and since we can have you know problem we have maximum K minus 1 regression variable. And the SS residual it decreases as the number of regression variables increases, so maximum number of regression variable possible is K minus 1. So, when this P is equal to K SS residual has the minimum value and hence the RP square, as you know will have the maximum value, so what we do here is that, we compute this the value of RP square.

So, basically fist we compute R 1 square this R 1 square is the case when the number of regression is equal to 0. So, R 1 square means this will have, so P call to 1; that means, P minus 1 equal to 0 the number of regressors in the model is equal to 0 that is the model if you consider the model y equal to beta naught plus epsilons. So, this is the model you know involving know regression variable, and it is not difficult to observe you know prove that, when we have this model with no regression variable.

Then the coefficient of multiple determinations is going to be equal to 0, next will be computing R 2 square given a set of data. So, R 2 square you know basically here P minus 1, so this is P, so P minus 1 is equal to 1, so this 1 is R 2 square is associated to the model y equal to beta naught plus beta 1 x 1 plus epsilons. So, this is R 2 square is for the model with one regresses.

(Refer Slide Time: 24:36)

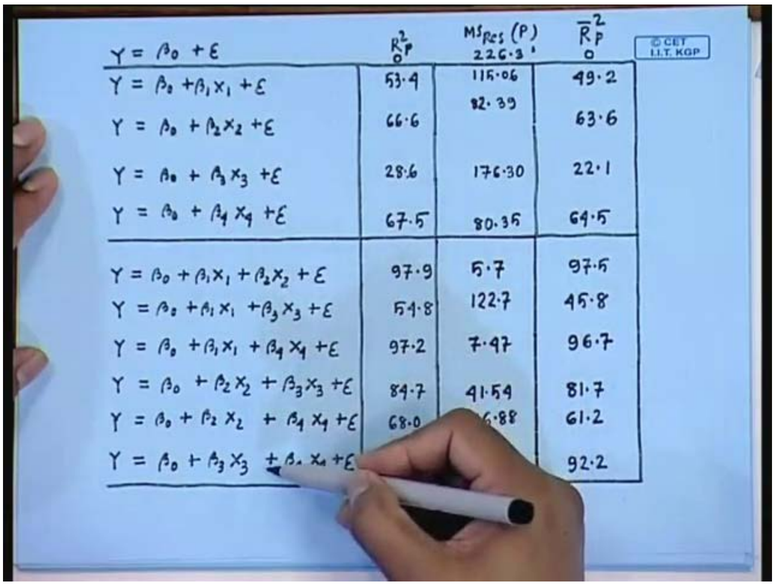

THE HALD CEMENT DATA

X_1	X_2	X_3	X_4	Y
7	26	6	60	78·5
1	29	15	52	74·3
11	56	8	20	104·3
11	31	8	47	87·6
7	52	6	33	95·9
11	55	9	22	109·2
3	71	17	6	102·7
1	31	22	44	72·5
2	54	18	22	93·1
21	47	4	26	115·9
1	40	23	34	83·8
11	66	9	12	113·3
10	68	8	12	109·4

Two illustrate all this things first I consider one example, while this is quit famous data this is called the HALD cement data. Here we have one response variable Y, and we have 4 regression variable X 1, X 2, X 3, and X 4 and we have 13 observations correspond to the response variable, and the regress variables well. Now, you know here we have four regression variables, and you may think that all four regress variables or not significant to explain the variability in while.

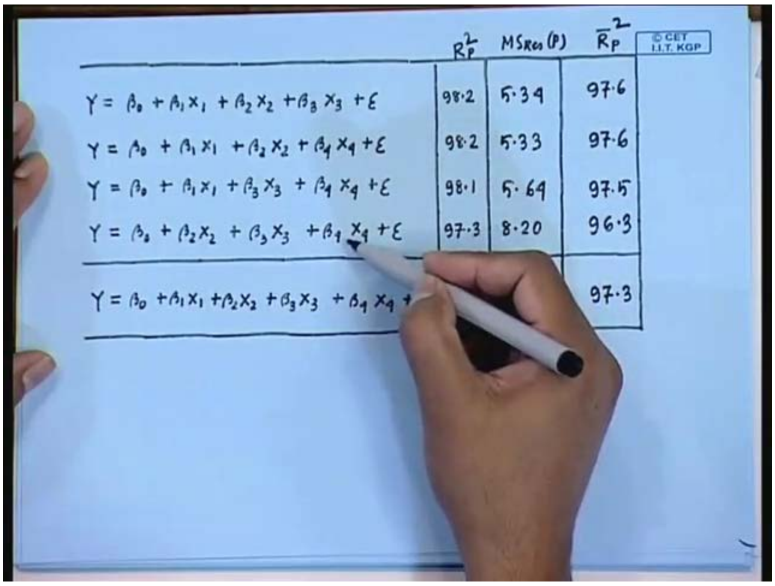
Some of them might be you know irrelevant and with the some variables can be removed from the model without affecting the model predicting power well. So, for that you know we need to select the regression variables, which regression variables are best to explain the variability in the responsible variables why that is the whole purpose of this lecture. Let me you know let me explain the all possible regression situation here using this example.

(Refer Slide Time: 26:42)



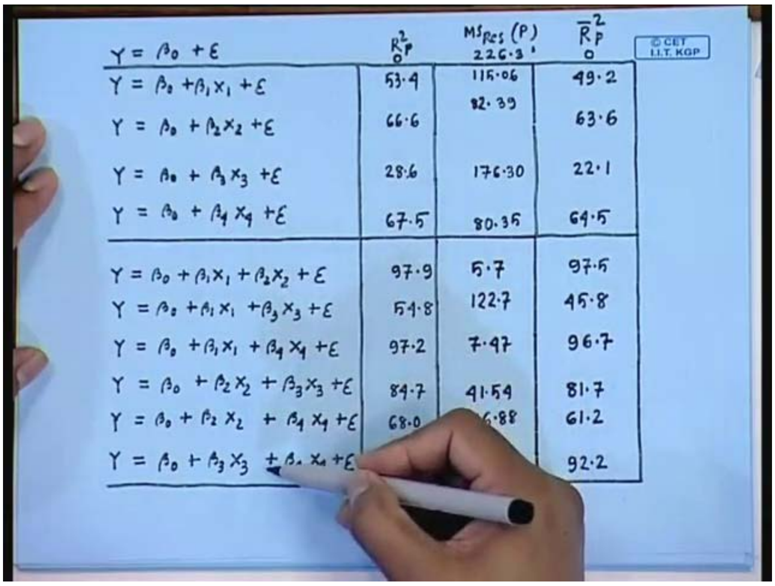
So, there are four regression variables, so these are the possible model this are the possible models with one regressors, these are the possible models with two regressors.

(Refer Slide Time: 26:53)



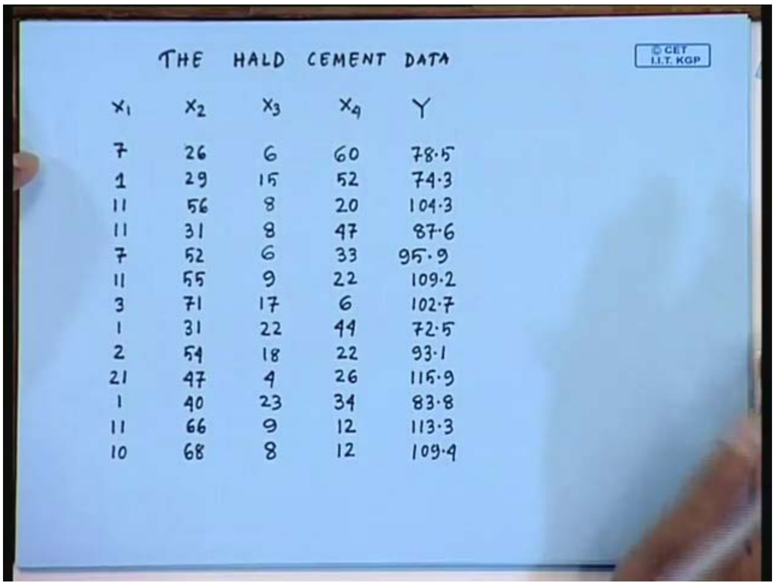
And these are the possible models with three regressors variables, and this is the model with four regressors variable.

(Refer Slide Time: 27:08)



Now, what we need to do is that, we need to feet each of them, and once you have the fitted equation or fitted model for this type of you know for involving x 1, you can compute the SS residual, SS total and from there you can compute the coefficient of multiple determination. Let me you know feet at least one equation for example, you know I will feet this equation.

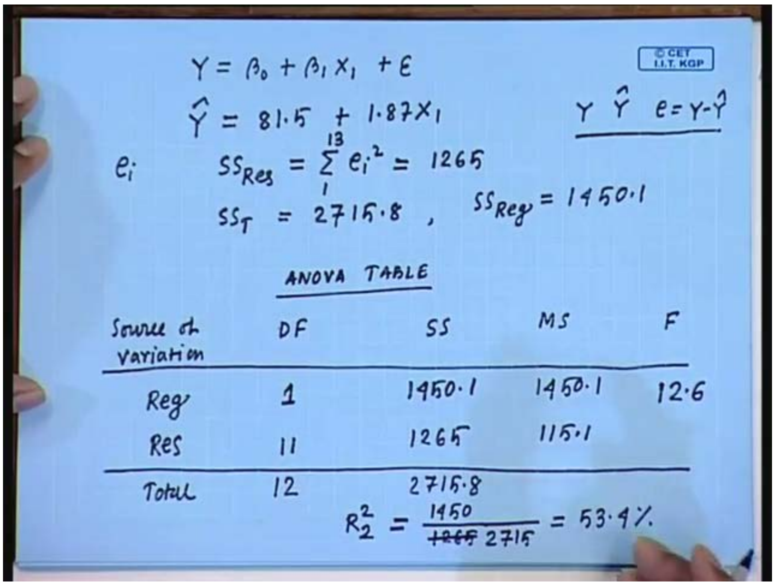
(Refer Slide Time: 27:48)


THE HALD CEMENT DATA

x_1	x_2	x_3	x_4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

So, I have this data I will try to feet a model between model of the firm Y equal to beta naught beta 1 X 1 plus epsilon right.

(Refer Slide Time: 28:12)



So, I will try to feet a model firm Y equal to beta naught plus beta 1 X 1 plus epsilon for that you know that HALD cement data I am not going into the detail of this looks like simple linear regression model. So, you know how to find beta naught hat, so you consider only the data corresponds to the response variable, and the data corresponds to the first regression variable X 1. And you know how to feed this model, fitting this model means you know they have to find the least of square estimate of beta naught and beta 1 hat.

So, the feted equation you can check that fitted equation is Y hat equal to 81.5 plus 1.87 X 1, so this is the feted a question. So, once you have the feted equation you can compute the residual e i, and once you have know e 1, e 2 up to e 13 you can compute the SS residual. So, SS residual is going to be e i square from 1 to 13, you just check you know this is equal to 1265 well, so you have the feted value, you have the original observation. So, from there you can get e which is equal to Y minus Y hat, so you know all this things.

And the SS total is equal to for this data it is 2715.8, and hence the SS regression is equal to 1450. So, I am just trying to give you some idea you know given a problem with four regressors or five regressors how to apply this all possible regression approach, now we can have the ANOVA table. So, ANOVA table for this problem I mean for this model is know you will write the source of variation degree of freedom SS, MS and the F statistic variation due to the regression model, total variation, the part remaining and explained residual.

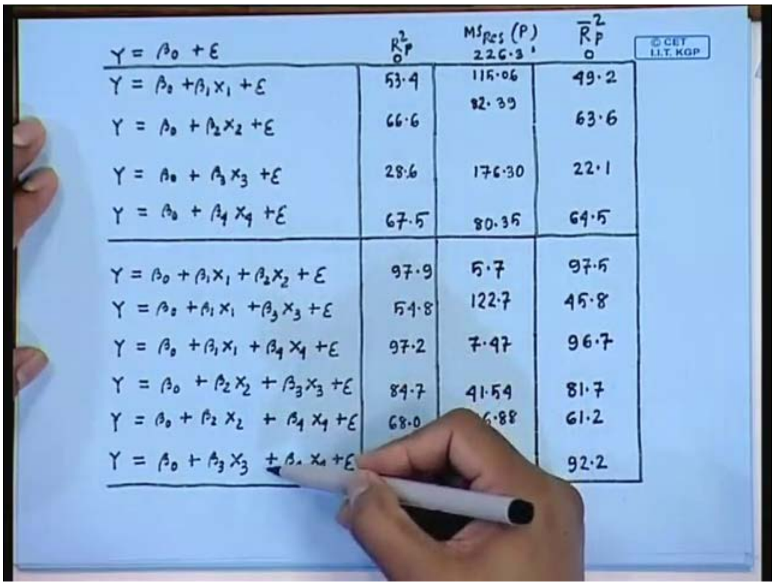
The total degree of freedom here is equal to 12 because there are 13 observation, now the SS residual you know here you have two unknown parameters. So, basically you will be getting two normal questions and; that means, there are two constant on the residuals, so the residual degree of freedom is equal to 13 minus 2 because of the two unknown parameters in the model. So, the SS residual has degree of freedom 11 and the regression degree of freedom is equal to 1 and we have the SS regression value is 1450.1, residual is 1265, in the total is 2715.8 right.

And MS value is 1450.1, and the MS value here is you know this is 115.1, and the F value is equal to 12.6 well. So, what I want to say here is that once, so this ANOVA table is associate with the model Y equal to beta naught beta 1 X 1 plus epsilon. Similarly you have to fit the other four models, involving one regression variable; that means, Y equal to beta naught plus beta 2 X 2 plus epsilon, so for that model you will get another ANOVA table.

Similarly you fit Y equal to beta naught plus beta 3 X 3 plus epsilon, you will get another ANOVA table Y equal to beta naught plus beta 4 X 4 plus epsilon you will have the ANOVA table associated with that model. So, basically you know there will be 16 possible regression models, and for each of them you will have you have to feat the model, you have to find out the associated ANOVA table for your convenience. Of course, you can use you know computer or some software package like SAS and S plus to do this job.

And then once you have you know all this ANOVA tables are the SS residual value is T value for every model you can compute the coefficient of multiple determination. So, here the coefficient of multiple determination R square and this is 2 or P is equal to 2 because there are 2 unknown parameters and here R square is equal to R 2 square is equal to SS regression, which is equal to 1450 by 1 to 65 by SS T which is equal to 2715, this is equal to 53.4 percent. So, here you know this model is not that good, because it explained the model involving the regression variable explains, this explains only 53 percents of the total variability in the response variable well.

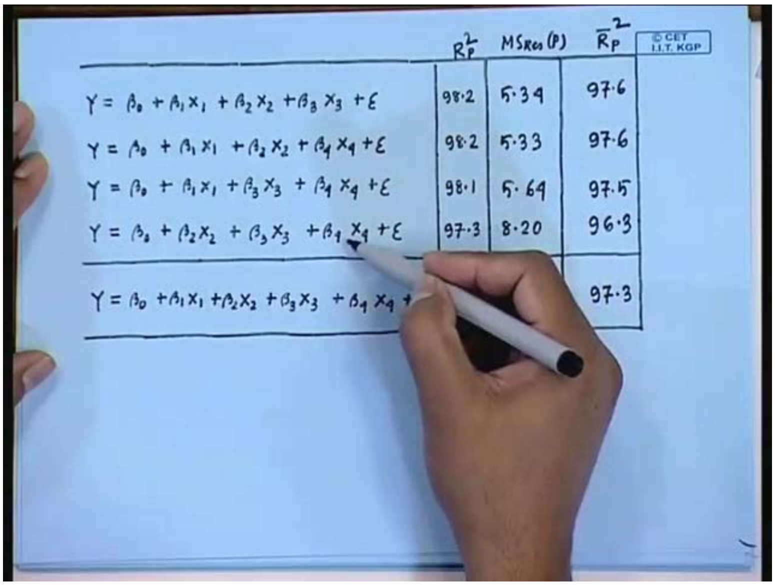
(Refer Slide Time: 36:45)



So, what I want to say now that look at this table here now, we have computed the coefficient of multiple determination for this model that is 53.4. Similarly, you fit this model, this model is also involving one regression variable and that is x 2 you find out the corresponding ANOVA table, and then you compute R square value. So, this is the R square coefficient of determination associated with this model and similarly you do for all the models here also you do for all the models.

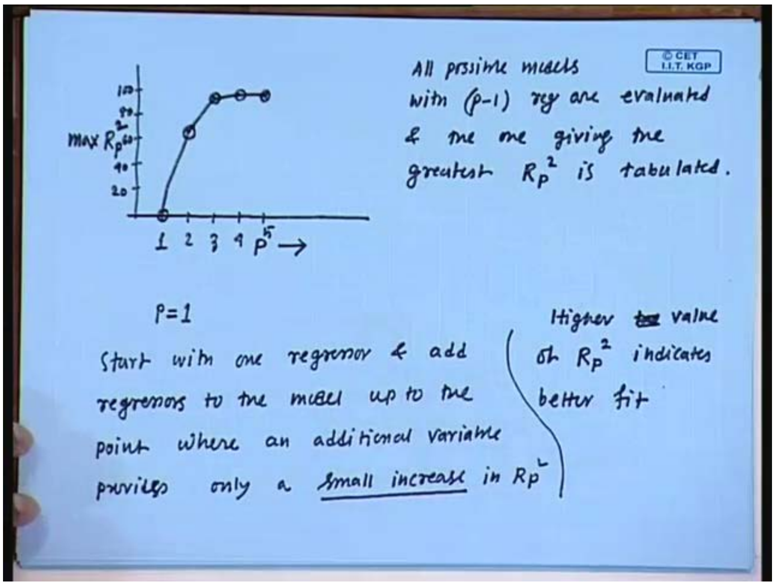
Here you can see that you know this model particularly it is a good one, this one involving x 1 and x 2 and the coefficient of multiple determination here is 97.9 percent; that means, which is maximum in this class. So, among the two variable among the regression equation, involving two variables this one is best this is; that means, Y equal to beta naught plus beta 1 X 1 plus beta 2 X 2 plus epsilon. Because, you know almost 98 percent of the total variability in the response variable has been explained by this model well.

(Refer Slide Time: 38:35)



So, similarly we have to you know this is really active job, you know here you have to estimate all the models involving three regressors. And you compute the R square value all the models, and this is the full model which involves all the four regressors, and the coefficient of determination is 98.2 well.

(Refer Slide Time: 39:15)



Now, what you want to do is that we want to draw a graph, the number of regressors variable P or basically P is the number of unknown parameters in the model. Along the x axis and maximum RP square, along the y axis I hope you know you have observe that the higher the value of RP square, better the model is are the higher value of RP square indicates better fit. So, what I want to mean is that out of all this six model, which involves two regressive variable this one is the best.

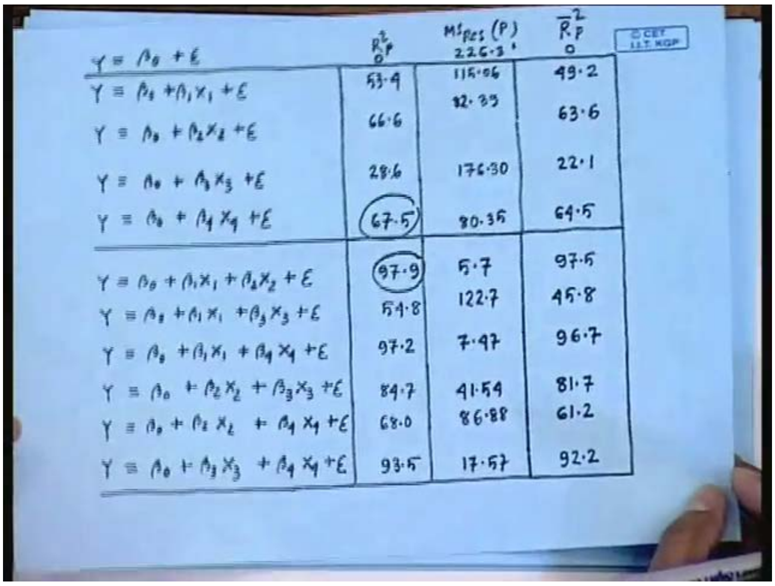
Out of all this four models involving one regress variable, this one is the best because this model as the maximum coefficient of determination well. So, what we do in this graph is that, here all possible models with P minus 1 regressors are evaluate using the criteria you know coefficient multiple determination, and the one giving the greatest RP square is tabulated. So, let me take this is my equal to 1, P equal to 2, P 3, P 4, P 5.

Now, when P equal to 1; that means, there is only one unknown in the model; that means, P equal to 1 means P minus 1 equal to 0; that means, there is no regressors in the model, and the RP square value is equal to 0 well. So, here the maximum RP square is equal to 0, now for P equal to 2, P equal to 2 means the number of regressors in the model is equal to 1.

So, out of this four model the maximum is 67.5, so will tabulated this one 67.5, suppose you know this is 20, 30, 40, 50, 60 may be 20 and then 40, 60, 80, 100. So, 67.5 we can keep it here, P equal to 3, P equal to 3 means number of regressors in the models is 2, and the maximum one is 97.9. So, will plot this 197.9, so for 3 it is almost here now for P equal to 4; that means, the number of regressors in the model is equal to 3 in the maximum is 98.2. So, for 4 it is 98.2 and for P equal to 5 means there are 4 regressors in the models, and the coefficient of determination value is 98.2 again.

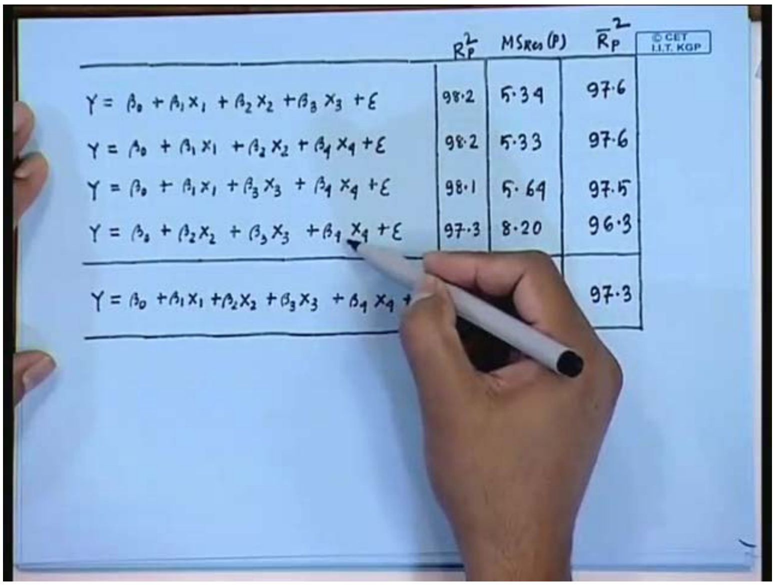
So, will plot 98.2 here well, so what it is suggest is that, the algorithm is like that you know you start with one regressors, and add regressors to the model up to the point where an additional variable provides only a small increase in RP square. So, best on this topic criteria, you can this small increase means there is no specific value of this model what you mean by small increase.

(Refer Slide Time: 46:55)



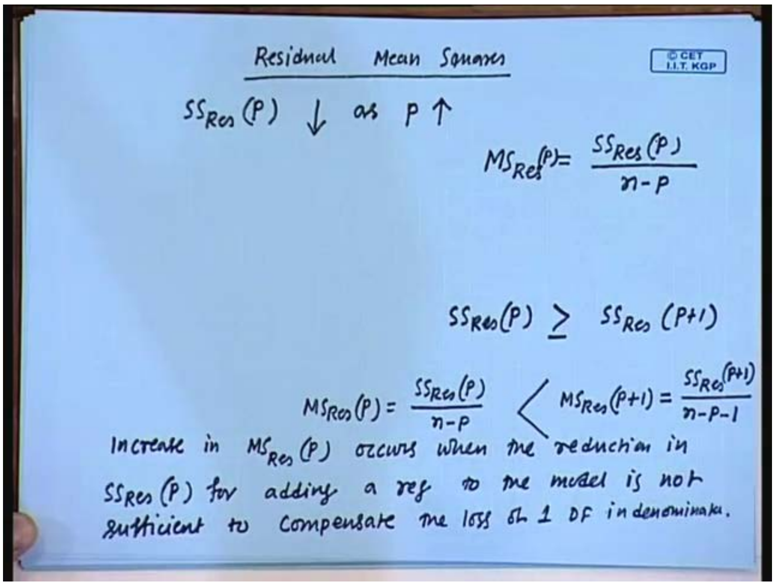
So, either you know this model with two variable it has coefficient of determination value of 97.9, which is close to 98 percent of the variability is explained by these two regression variable.

(Refer Slide Time: 47:10)



If you go for the three variable model, then this one is best or also this one is also the I mean same multiple regression model. So, clearly you know you do not need to go for the four variable model, either you choose the three variable model which is you know beta 1 Y equal to beta naught plus beta 1 X 1 plus beta 2 X 2 plus beta 3 X 3, either you go for this model or you go for this model. And according to the coefficient of multiple determination criteria, this one is also not bad you know this is the model with two regressors well. So, this is how we evaluate the different possible basically all possible models using some criteria. So, we talked about one criteria that is coefficient of multiple determination, and the next will more for the MS residual criteria.

(Refer Slide Time: 48:27)



Residual mean square well, so what you know is that SS residual P by P I mean you know this is the SS residual for the model, which for the model with P minus 1 regressors. When k minus 1 is the total number of regressive variables, we know that this one decreases, as the number of regression variable increases, and here we are talking about MS residual, which is equal to SS residual P. Let me denoted by P also MS residual P by n minus P, n minus P is the degree of freedom for the associated model.

Either degree of residual degree of freedom for the associated model, and here one thing you know I want to mention that, you know for SS residual decreases as P increases, but this is not true for MS residual I mean MS residual may increase with P. So, reason behind this one is that what I want to say here, let me write MS residual P which is equal to SS residual P by n minus P. And also let me write MS residual P plus 1, which is equal to SS residual P plus 1 by n minus P minus 1.

We know that, SS residual P is greater then what you call to SS residual P plus 1 because SS residual decreases has P increases. But, the same thing is not true for MS residual, here this could be larger than the MS residual P, the reason is this you know the increase in MS residual P occurs, I mean this may be I mean larger this occurs when the reduction in SS residual P for adding regressors to the model is not sufficient to compensate the loss of 1 degree of freedom, in denominator.

Of course, what I want to say here is that know this one is of course, smaller than this one, but if you add and irrelevant regressors in the model this will decrease. But, the reduction here for adding one more regression in the model, the reduction in SS residual is if it is not sufficient to compensate you know 1 degree of freedom loss here, than only it increases. So, if you the newly added regressors variable is not relevant for the response variable, are not relevant for the model.

Then only you know the reduction in SS residual for adding this irrelevant regressors the model is not sufficient to compensate the 1 degree of freedom loss in the model, then only MS residual increases well. So, we learned how to evaluate you know all possible models using the MS residual criteria, in the next class well. So, will continues this criteria MS residual in the next class.

Thank you for your attention.