

Quality Control and Improvement with MINITAB
Prof. Indrajit Mukherjee
Shailesh J. Mehta School of Management
Indian Institute of Technology, Bombay

Lecture - 26
Best Subset Regression, Multicollinearity

Hello and welcome everyone to this course on Quality Control and Improvement with MINITAB. We are in session 26, and I am Professor Indrajit Mukherjee from Shailesh J Mehta School of Management, IIT Bombay.

So, in last session, we are discussing about multiple regression and how to select the variables. You see there are confusions which arises when we have multiple x which is regressed with a y, single y. So, we are not sure which variable should be considered, which is not to be considered. So, some examples we have taken last time. So, let us try to see whether we can resolve that problem and dilemmas that we are facing.

(Refer Slide Time: 00:50)

Quality Control and Improvement using MINITAB

Example

The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature (x_1), the number of days in the month (x_2), the average product purity (x_3), and the tons of product produced (x_4).

The past year's historical data are available and are presented in the given table. Obtain a least square regression model.

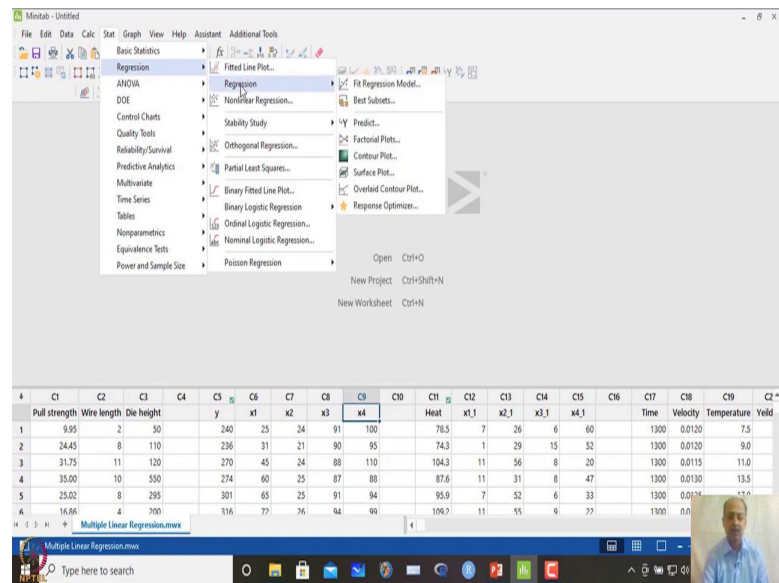
Y	X_1	X_2	X_3	X_4
240	25	24	91	100
236	31	21	90	95
270	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

Data Source: Montgomery, D. C. (2005).
Applied statistics and probab
engineers. John Wiley & Son

Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay

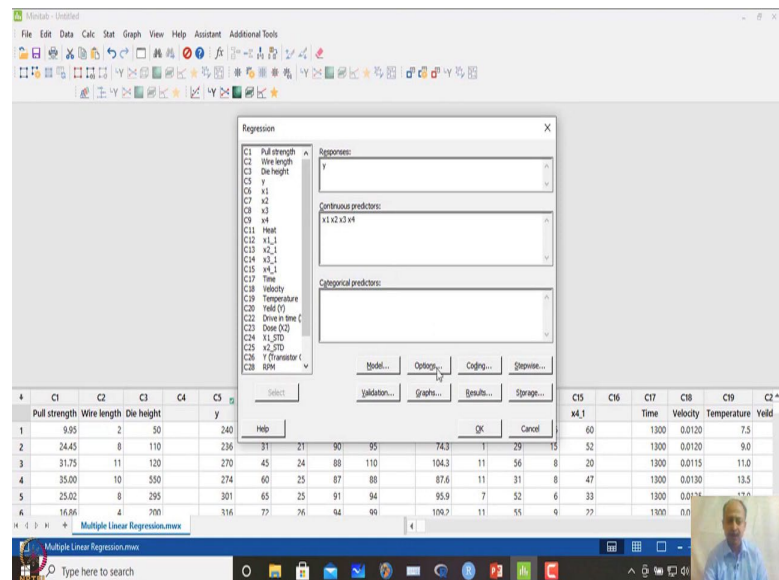
So, this was one of the problem that we are dealing with at the time point, that electrical power consumption is monitored over here. And this may be related to variable X_1 , X_2 , X_3 , and X_4 and the details are given on the left hand side of your screen in this power point. And we want to see which is the best model that explains y with respect to given x, set of x like that.

(Refer Slide Time: 01:20)

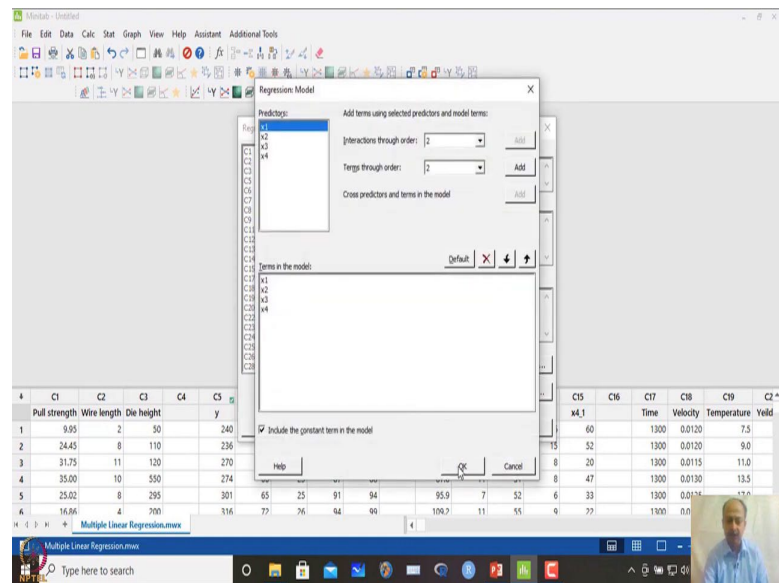


So, what we have done is that we have gone to MINITAB, and then what we have seen is that these are the variables this from C5 to C9, C5 is the y and C6 to C9 is the 4 variables. So, what we have done is that stat we have gone to regression, and we have gone to regression, and then we have used fit regression models.

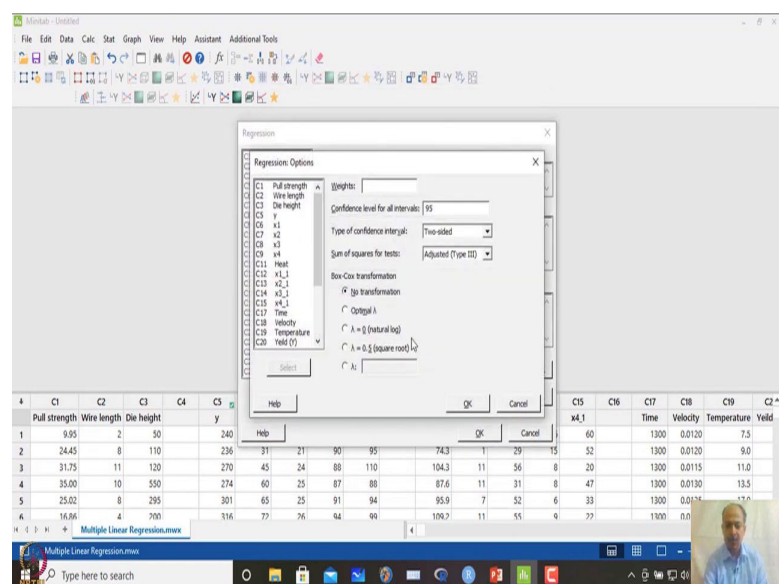
(Refer Slide Time: 01:40)



(Refer Slide Time: 01:49)

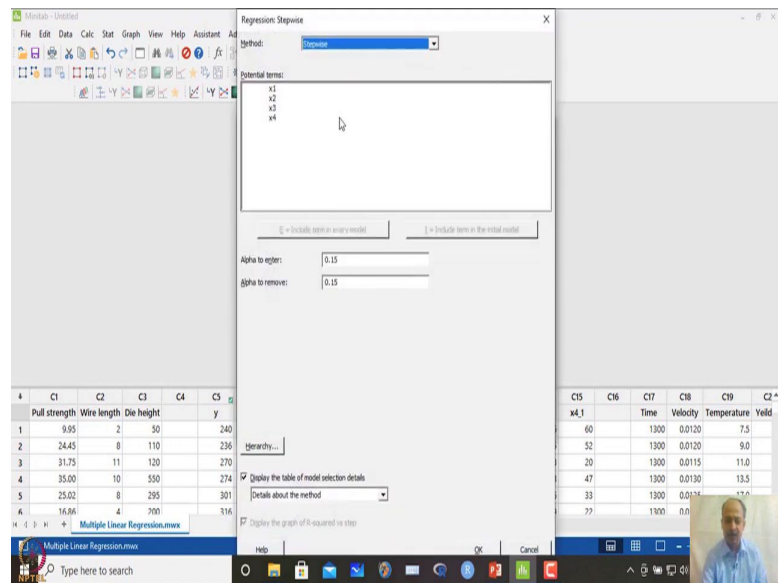


(Refer Slide Time: 01:55)

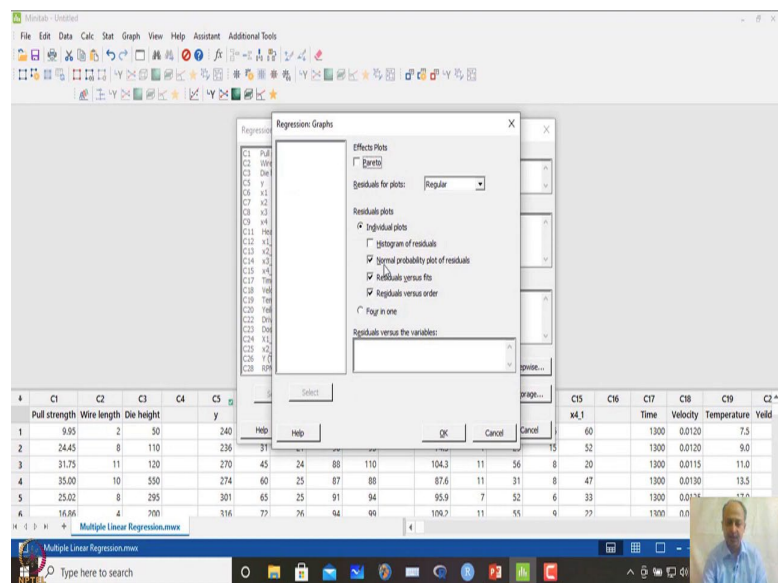


And then what we have done is that we have selected y and the set of x conditions that is X1, X2, X3, and X4. And in models what we have done is that all the variables we have considered include a constant term. And options we have not given any transformation over here.

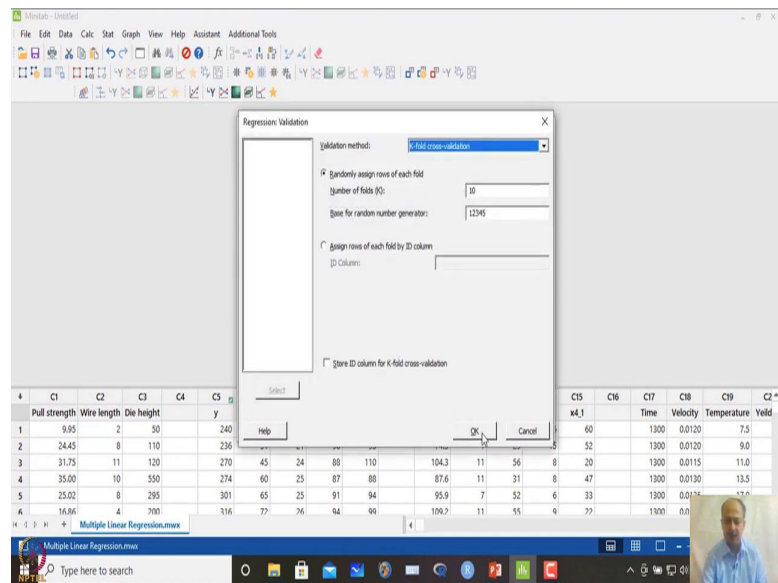
(Refer Slide Time: 02:00)



(Refer Slide Time: 02:07)

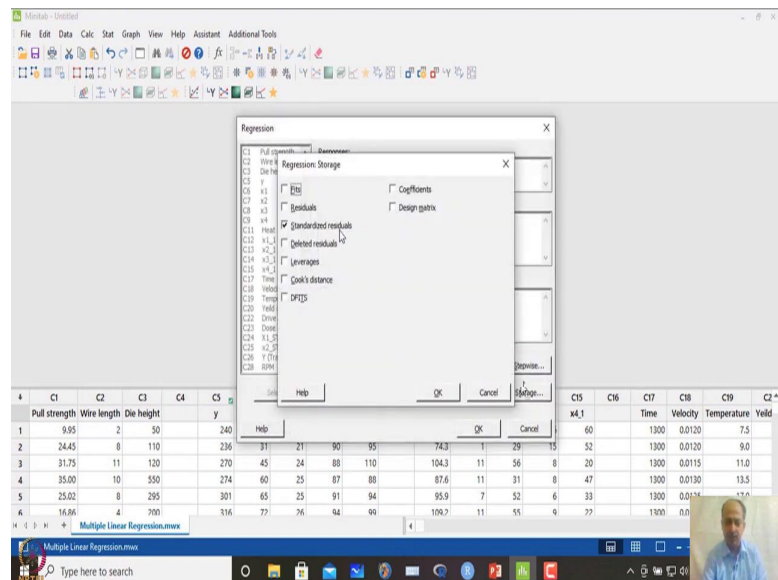


(Refer Slide Time: 02:16)



And in this case first we have not done not given any let us say stepwise regression over here. And we may see graphs over here, normal probability plot, residual plots and order plots like that. So, this is possible. And validation tenfold cross-validation that we have discussed last time also we can put over here.

(Refer Slide Time: 02:21)



(Refer Slide Time: 02:25)

Regression Analysis: y versus x1, x2, x3, x4

Method
Cross-validation 10-fold

Regression Equation
 $y = -123 + 0.797 x_1 + 7.52 x_2 + 2.48 x_3 - 0.481 x_4$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-123	157	-0.78	0.459	
x1	0.797	0.279	2.71	0.030	2.32
x2	7.52	4.01	1.87	0.103	2.16

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1,1	x2,1	x3,1	x4,1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120		7.5
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120		9.0
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115		11.0
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130		13.5
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0130		13.0
6	16.86	4	700		316	77	26	94	99		106.7	11	55	9	77		1300	0.0130		13.0

(Refer Slide Time: 02:32)

Regression Analysis: y versus x1, x2, x3, x4

Method
Cross-validation 10-fold

Regression Equation
 $y = -123 + 0.797 x_1 + 7.52 x_2 + 2.48 x_3 - 0.481 x_4$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-123	157	-0.78	0.459	
x1	0.797	0.279	2.71	0.030	2.32
x2	7.52	4.01	1.87	0.103	2.16
x3	2.48	1.81	1.37	0.212	1.34
x4	-0.481	0.355	-0.87	0.415	1.01

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	10-fold S	10-fold R-sq
11.7666	85.20%	76.75%	52.48%	16.8035	48.45%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Constant	1	15111.0	15111.0	0.00	0.999
x1	1	11.8	11.8	0.03	0.861
x2	1	11.8	11.8	0.03	0.861
x3	1	11.8	11.8	0.03	0.861
x4	1	11.8	11.8	0.03	0.861
Error	5	11.8	2.36		
Total	11	23.6			

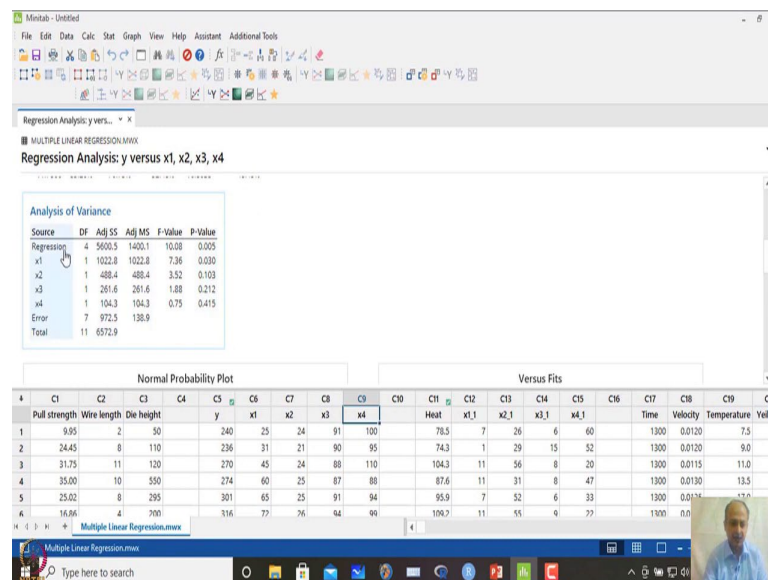
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1,1	x2,1	x3,1	x4,1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120		7.5
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120		9.0
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115		11.0
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130		13.5
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0130		13.0
6	16.86	4	700		316	77	26	94	99		106.7	11	55	9	77		1300	0.0130		13.0

And then storage I want to store that the standardized residual let us say and click ok. So, let us try to see what are the results that we are getting. So, MINITAB gives you automatically a regression equation over here, but what is surprising over here you see that although X1 is the only variable which is significant over here; that means, less than 0.05 and others are not significant terms over here.

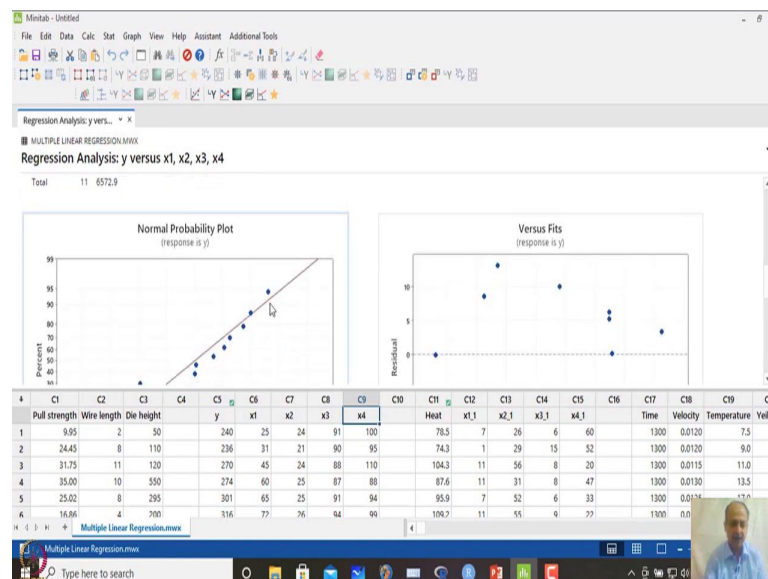
So, only X1 is showing significance, others are not showing significance over here. And also you see the R square value adjusted R square is about 76 and tenfold R square this

cross-validation is around 48. So, when there is a difference between these two, significant difference that exist over here there must be something going wrong over here. And maybe model fitting is not correct or we have over-fitted the model basically, ok.

(Refer Slide Time: 03:14)



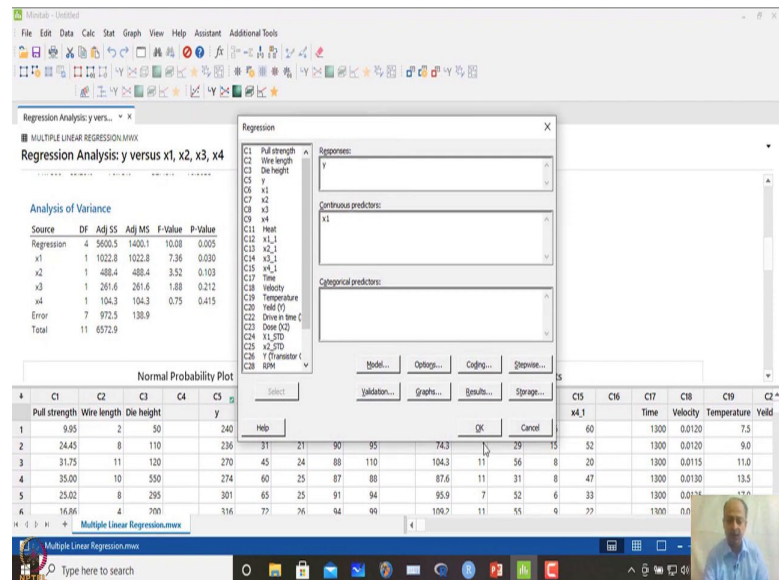
(Refer Slide Time: 03:18)



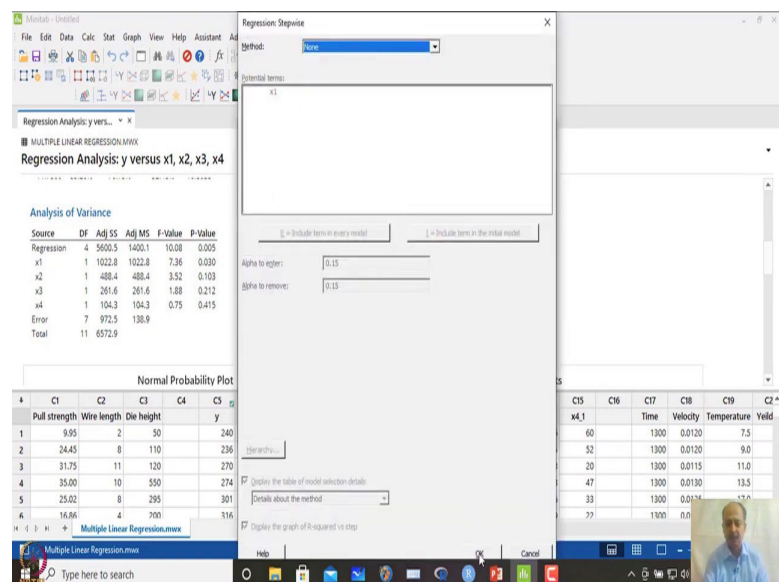
So, in this scenario when we are not sure which variable will go in which variable will go out. And analysis of variance also confirm that one of the variable X1 is significant others are not significant what we can do is that, we can remove all the variables all the

X2, X3, X4s like that and retain only X1 like that only retain X1 and then we can see. So what will happen? So, this is trial and error methods that I am showing over here without going into our stepwise regression like that.

(Refer Slide Time: 03:36)



(Refer Slide Time: 03:42)



(Refer Slide Time: 03:47)

Regression Analysis: y versus x1

Method
Cross-validation 10-fold

Regression Equation
 $y = 219.4 + 1.011 x_1$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	219.4	14.3	15.38	0.000	
x1	1.011	0.238	4.25	0.002	1.00

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1,1	x2,1	x3,1	x4,1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120		7.5
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120		9.0
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115		11.0
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130		13.5
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0130		17.0
6	16.86	4	700		316	77	26	94	99		106.7	11	55	9	77		1300	0.0		0.0

(Refer Slide Time: 03:49)

Regression Analysis: y versus x1

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	219.4	14.3	15.38	0.000	
x1	1.011	0.238	4.25	0.002	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	10-fold S	10-fold R-sq
15.2955	64.61%	60.85%	50.24%	16.5254	50.14%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4233.41	4233.41	18.10	0.002
x1	1	4233.41	4233.41	18.10	0.002
Error	10	2416.61	241.66		

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1,1	x2,1	x3,1	x4,1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120		7.5
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120		9.0
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115		11.0
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130		13.5
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0130		17.0
6	16.86	4	700		316	77	26	94	99		106.7	11	55	9	77		1300	0.0		0.0

So, I can remove this X1 which are not significant over here and we could have regressed this one and a stepwise regression we have done none over here. So, in this case I can click ok and see what is the performance of that model. And what we are seeing is that, ok, X1 is significant that is shown over here regression coefficient and the equation is also given and, but the explained variability is very less R square adjusted is around 60.85 and tenfold cross-validation is about 50 percent.

Although, now, there is somewhat match between cross-validation and also R square adjusted somewhat close we can we can assume over here. But still I feel that there could be something more which can be done on this, ok.

(Refer Slide Time: 04:12)

Regression Analysis: y versus x1

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4233.41	4233.41	18.10	0.002
x1	1	4233.41	4233.41	18.10	0.002
Error	10	2339.51	233.95		
Lack-of-Fit	9	2337.51	259.72	129.88	0.068
Pure Error	1	2.00	2.00		
Total	11	6572.92			

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4233.41	4233.41	18.10	0.002
x1	1	4233.41	4233.41	18.10	0.002
Error	10	2339.51	233.95		
Lack-of-Fit	9	2337.51	259.72	129.88	0.068
Pure Error	1	2.00	2.00		
Total	11	6572.92			

Regression

Responses: y

Continuous predictors: x1, x2, x3, x4

Categorical predictors:

Model... Options... Coding... Storage... Validation... Graphs... Results... Storage...

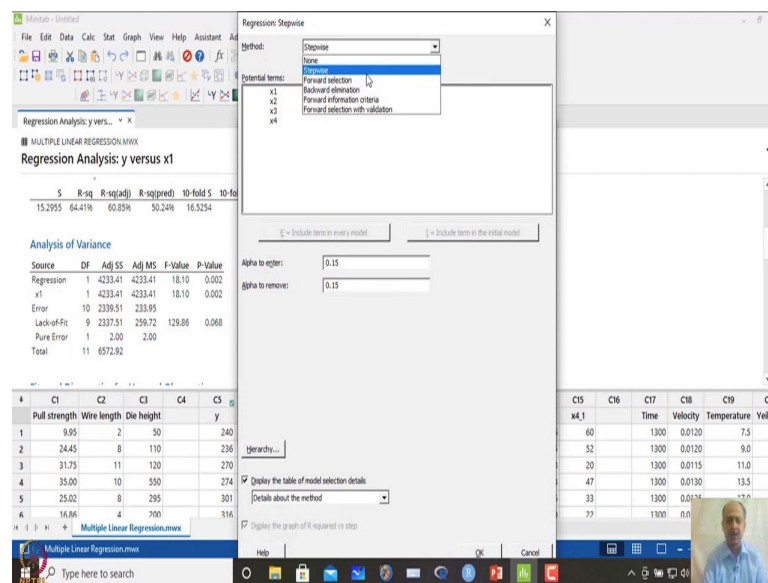
Help

Multiple Linear Regression.mnx

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Pull strength	Wire length	Die height		y							
1	9.95	2	50	240	31	21	90	95	74.3	1	29
2	24.45	8	110	236	45	24	88	110	104.3	11	56
3	31.75	11	120	270	60	25	87	88	87.6	11	31
4	35.00	10	550	274	60	25	91	94	95.9	7	52
5	25.02	8	295	301	65	25	91	94	95.9	7	52
6	16.86	4	700	316	77	36	94	99	104.7	11	55

So, lack of fit test also shows that we are above 0.05, so in this case there is no sign of lack of fit. So, then and to avoid all this confusion what we can do is that directly we can use that stepwise regression over here. So, what you have to do is that fit regression model and go to instead of this variable you select all the variables that is X1, X2, X3, and X4 over here.

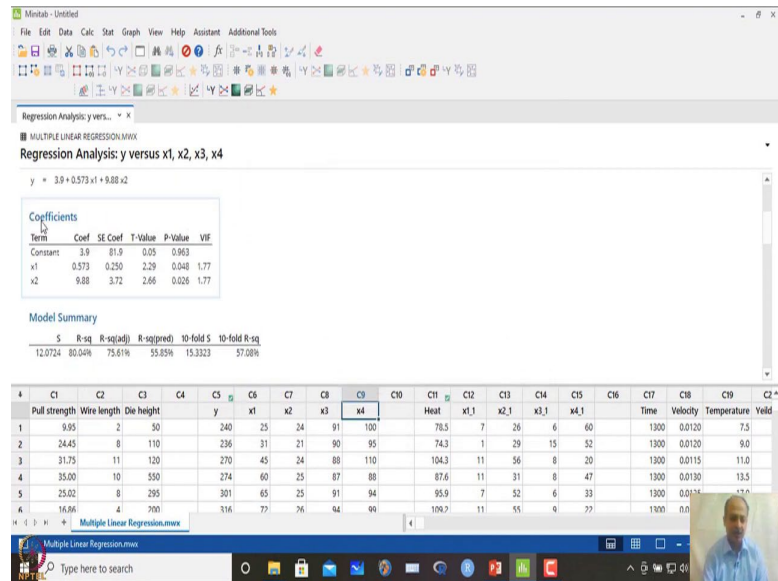
(Refer Slide Time: 04:42)



And in stepwise options what you do is that right stepwise over here, and there will be alpha to enter alpha to remove this to you can keep, so this methodology works, this stepwise regression works like adding variables and removing variables like that simultaneously. It will work in adding and removing variables which is significant which is not significant like that.

So, that way it works and the theories can be seen in any books on stepwise regression, you have details about this stepwise regression. There are other methods of stepwise regression which is also forward selection and backward elimination these are the other two methods. But stepwise include forward selection and backward elimination both of them. So, we prefer to use stepwise regression over here. So, in this case I click ok.

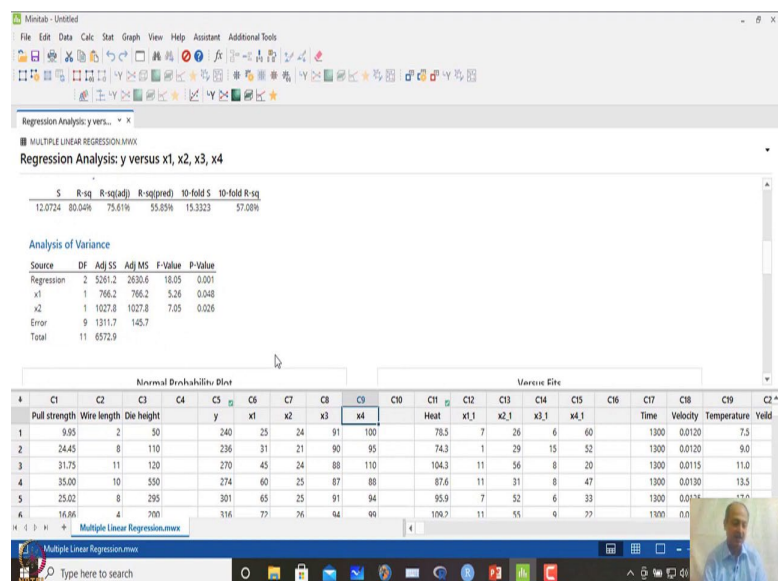
(Refer Slide Time: 05:29)



And then if you click ok over here what will happen is that you will find a scenario over here which is two variables are entered into the model that is X1. And you can see that X1 is also prominent over here 0.048.

And the other one is X2 is more prominent which is P value is 0.026 both are significant and less than 0.05 over here. And also R square adjusted has improved significantly from 60 to 75 over here and this tenfold cross-validation from 50 to 57 it has improved over here.

(Refer Slide Time: 05:58)



So, in this case, we can say that this may be the best model that we have considered over here and we should go we should go about go in implementing this one.

(Refer Slide Time: 06:14)

The screenshot shows the Minitab software interface. The 'Stat' menu is open, and the 'Regression' option is selected. The 'Model Summary' table is displayed, showing the following data:

S	R-sq	R-sq(adj)	R-sq(pred)	10-fold S
12.0724	80.04%	75.61%	55.85%	15.332

The 'Coefficients' table is also visible, showing the following data:

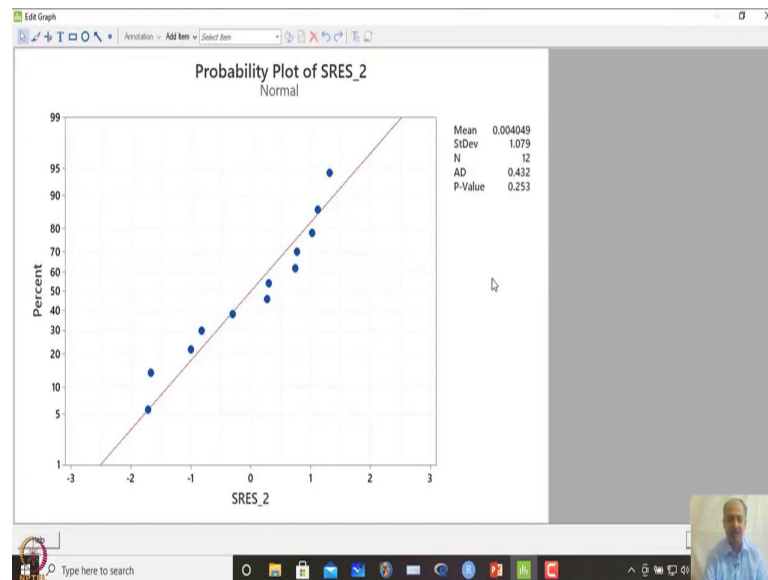
Term	Coef
Constant	3.9
x1	0.573
x2	9.88

(Refer Slide Time: 06:17)

The screenshot shows the Minitab software interface with the 'Normality Test' dialog box open. The 'Variable' field is set to 'RES_1'. The 'Tests for Normality' section is checked, and the 'Anderson-Darling' test is selected. The 'Model Summary' table is also visible, showing the following data:

S	R-sq	R-sq(adj)	R-sq(pred)	10-fold S	10-fold R-sq
12.0724	80.04%	75.61%	55.85%	15.3323	57.08%

(Refer Slide Time: 06:22)

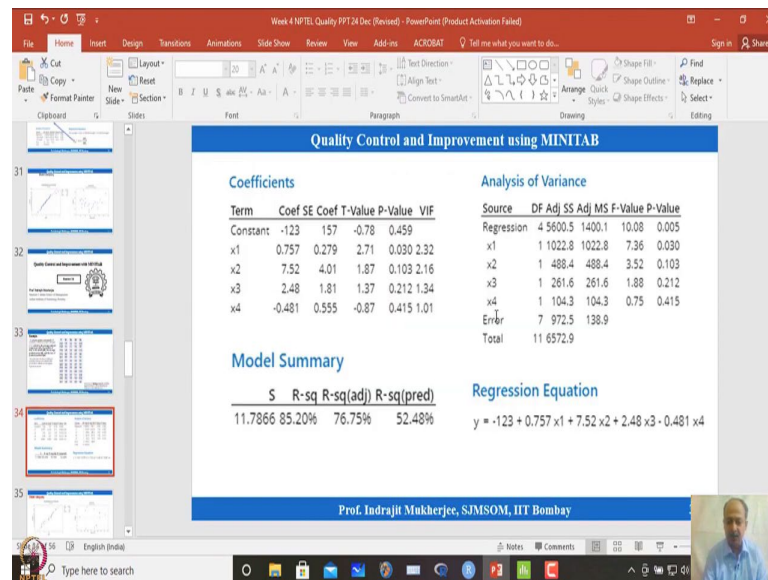


So, this may be this model that we have generated over here. Now, only thing is that residual whether it is normal distribution we can check by normality test over here. And the residual will be saved at the end values that we see over here residual 2 that is the final residual after we have run the models. And the residual also says that there is no abnormality in the normality assumptions that we are considering in the residuals over here.

0.253 is more than 0.05, so in this case residual also satisfies. So, there is no problem. So, if I consider X1 and X2 in the model with y. We are getting X1, X2 significant and also the model adequacy test are quite there is no significant deviation from that, ok.

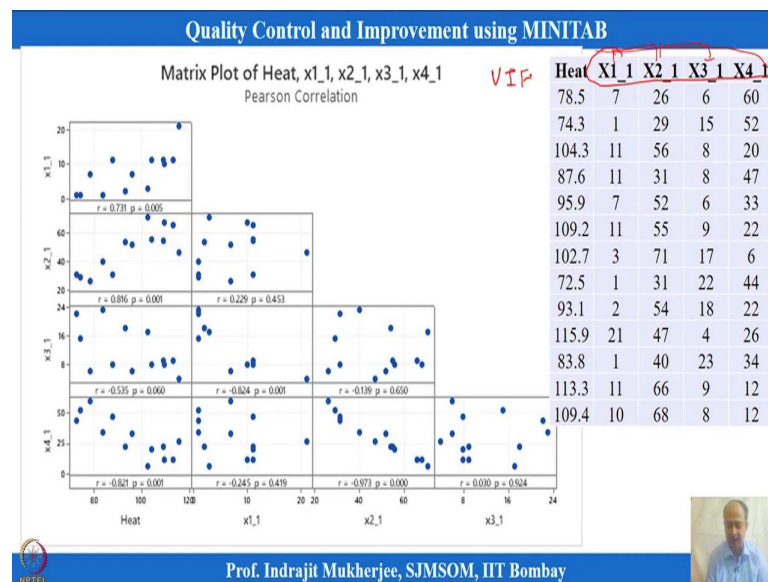
So, this is one of the scenario where we can use stepwise regression method. But for the second case if we consider this one that heat and X1, X2 again the confusion comes, again the confusion will arise because and we can see that one. So, what we will do is that I will take this second one where heat is the y characteristics and that has to be regressed with other variables over here.

(Refer Slide Time: 07:14)



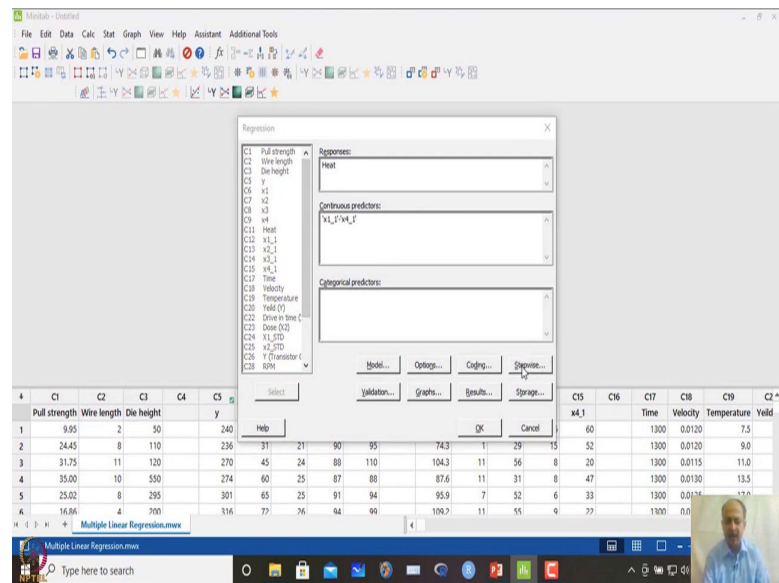
So, this is the first example that we have selected and these are the values that we have seen X1, X2, this is with X3.

(Refer Slide Time: 07:22)



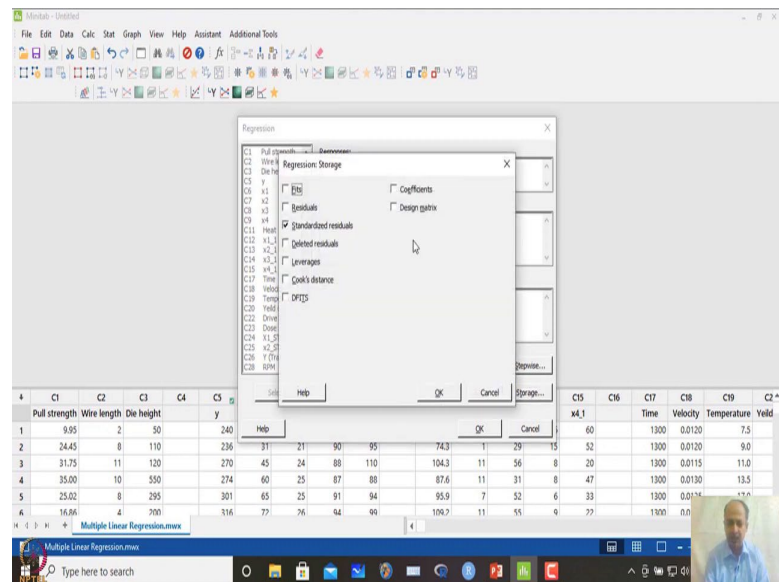
So this is now, this is the variables that we are talking about heat is with X1, X2, X3, and X4, 4 variables we are trying to regress.

(Refer Slide Time: 07:40)

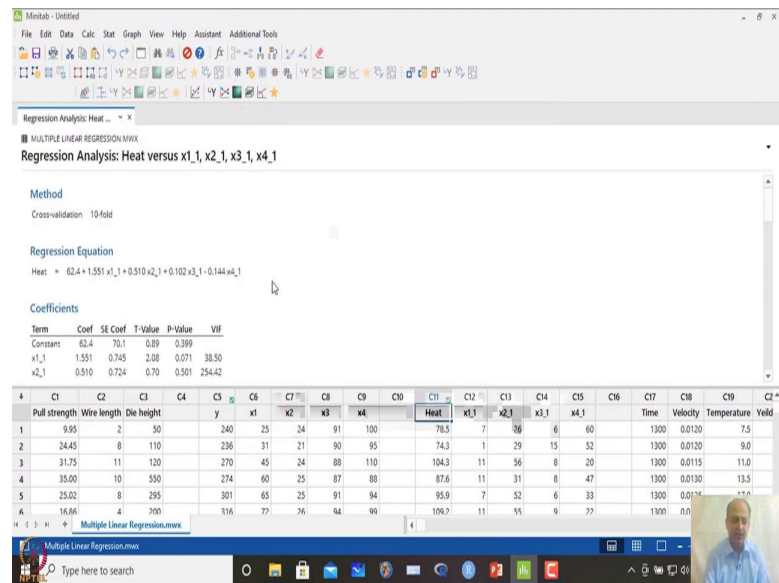


And then what we can do is that I can go to stat, and I can go to regression and then regression and fit regression models over here, and instead of this I give heat and then I give X1 to X4 variables I select this one. And I do not do stepwise regression at initially. So, I want to see what are the results.

(Refer Slide Time: 08:01)



(Refer Slide Time: 08:05)



Regression Analysis: Heat versus x1_1, x2_1, x3_1, x4_1

Method
Cross-validation 10-fold

Regression Equation
Heat = 62.4 + 1.551 x1_1 + 0.510 x2_1 + 0.102 x3_1 - 0.144 x4_1

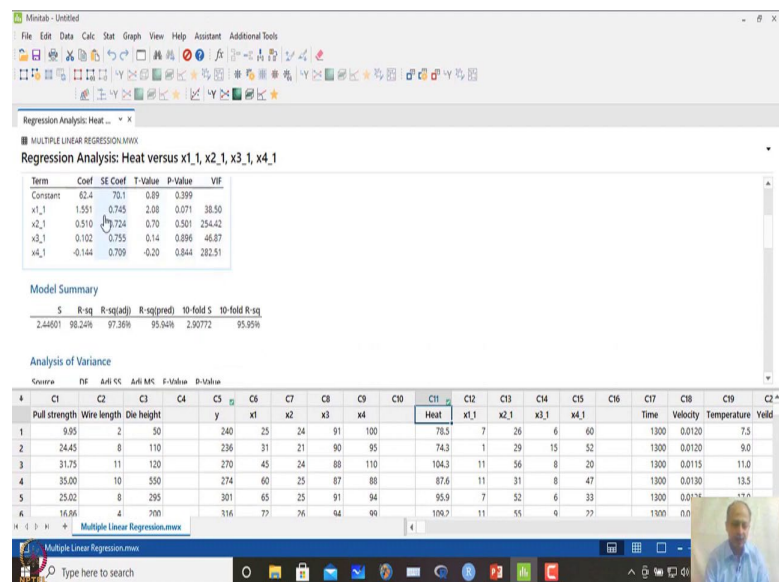
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	62.4	70.1	0.89	0.399	
x1_1	1.551	0.745	2.08	0.071	38.50
x2_1	0.510	0.724	0.70	0.501	254.42

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1_1	x2_1	x3_1	x4_1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120	7.5	
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120	9.0	
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115	11.0	
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130	13.5	
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0130	13.5	
6	16.86	4	700		316	77	26	94	99		106.7	11	55	9	77		1300	0.0130	13.5	

So, what will happen is that I have all the criteria. So, and I will store let us say the residuals over here and I click ok.

(Refer Slide Time: 08:08)



Regression Analysis: Heat versus x1_1, x2_1, x3_1, x4_1

Method
Cross-validation 10-fold

Regression Equation
Heat = 62.4 + 1.551 x1_1 + 0.510 x2_1 + 0.102 x3_1 - 0.144 x4_1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	62.4	70.1	0.89	0.399	
x1_1	1.551	0.745	2.08	0.071	38.50
x2_1	0.510	0.724	0.70	0.501	254.42
x3_1	0.102	0.755	0.14	0.896	46.87
x4_1	-0.144	0.709	-0.20	0.844	282.51

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	10-fold S	10-fold R-sq
2.44001	98.24%	97.36%	95.04%	2.90772	95.95%

Analysis of Variance

Source	DF	SS	Adj SS	Adj R-Sq	F-Value	P-Value
Regression	4	16.86	16.86	98.24%	106.7	0.000
Error	1	0.00	0.00			1.000
Total	5	16.86				

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1_1	x2_1	x3_1	x4_1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120	7.5	
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120	9.0	
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115	11.0	
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130	13.5	
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0130	13.5	
6	16.86	4	700		316	77	26	94	99		106.7	11	55	9	77		1300	0.0130	13.5	

I want to see what model says? So, over here what you see is that X1, X2, P values, if you observe P values none of the P values is significant, but R square adjusted is highly I am getting a very good model fit over here, 97 over here, and the tenfold cross-validation is also very good over here. So, something wrong is happening because none of the variable is significant. But I am able to predict very high predictability over here what

we are seeing like that. So, this equations means we can say that let us adopt this one, but we will not do that because there is another issue coming over here which is known as variation inflation factor that is, a variation inflation factor over here.

So, variation inflation factor indicates that whether there is a situation of multicollinearity in the data set. So, what is multicollinearity? We will try to explain. Multicollinearity in a sense it says that whenever the x are interrelated with each other let us say X1 with X2 or X2 with X3, and the correlation is very high that will influence the model and the model will not be correct and it will give you a bias judgment and the sign that you will get coefficient sign that you will get may interchange.

That means what it should be positive it is reflecting negative like that. So, that can happen over here. So, multicollinearity means there is a high significant relationship between the x variables over here. And this will be reflected by a index that is known as variation inflation factor, that is known as variation inflation factor.

(Refer Slide Time: 09:34)

Quality Control and Improvement using MINITAB

Best Subset Regression

Following table presents data concerning the heat evolved in calories per gram of cement as a function of four different chemical ingredients.

Heat	X1	X2	X3	X4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Response is Heat

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	1	1	1	1
1	67.5	64.5	56.0	138.7	8.9639				X
1	66.6	63.6	55.7	142.5	9.0771				X
2	97.9	97.4	96.5	2.7	2.4063				X X
2	97.2	96.7	95.5	5.5	2.7343				X X
3	98.2	97.6	96.9	3.0	2.3087				X X X
3	98.2	97.6	96.7	3.0	2.3121				X X X
4	98.2	97.4	95.9	5.0	2.4460				X X X X

Source: Montgomery, D. C., Peck, E. A., Vining, G. G. (2003). Introduction to linear regression analysis. John Wiley & Sons

Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay

(Refer Slide Time: 09:36)

Quality Control and Improvement using MINITAB

Multicollinearity

When there are **near-linear dependencies among the regressors**, the problem of multicollinearity is said to exist.

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

$X_1 \rightarrow$	1	(2)	(3)	4	5	6
$X_2 \rightarrow$	(2)	(4)	(6)	8	10	12

$$VIF_{X_1} = VIF_{X_2}$$

$$X_1 = f(X_2)$$

$$(R_{12})$$

$$X_2 = 2X_1$$

$$VIF_i = \frac{1}{1 - R_i^2}$$



Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay



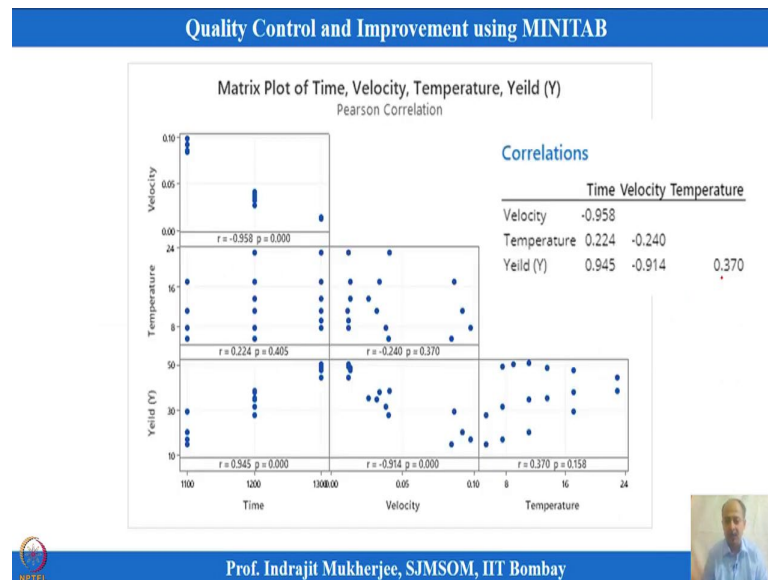
And this is what I mean to say over multicollinearity over here. So, you see the numbers of X_1 is given over here for a particular and X_2 is also simultaneously recorded. So, X_1 is 1, X_2 is 2 and when this is 2 this is 4 like this, this is 3 this is 6 like this. So, X_2 has a functional relationship with X_1 over here.

So that means, there is a high amount of correlation between the data set that I am having in X_1 and the data set that I am having in X_2 . So, I can calculate the variation inflation factor between this data set. So, variation inflation factor for X_1 , I can calculate the similarly for X_2 , I can calculate like that only two variables over here.

So, then I calculate the R_i index that is the coefficient of determination over here, so R_i index can be calculated, where X_1 let us say is regressed with a function of X_2 like this. And then the R values are indicated, so that is with one and two like that. So, this values will be indicated and that will be reported over here we can put that value and I can get the variation inflation factor.

So, variation inflation factor for the first variable X_1 will be same as variation inflation factor for because there is only two variables over here, say this one and this one like that; so, X_1 and X_2 if I am considering two variables over here.

(Refer Slide Time: 10:51)



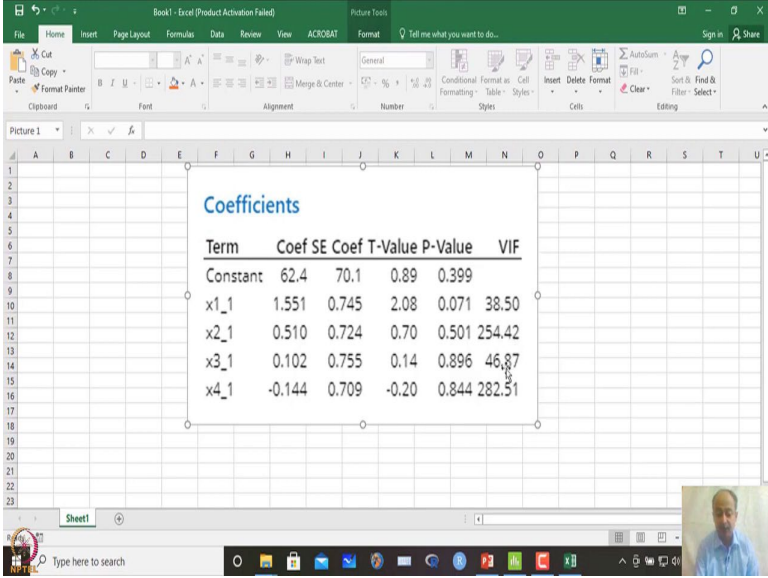
So, this variation inflation factor what would you see will lead to models; we will not get the best models out of this when multicollinearity exists, and then the beta coefficient estimation goes wrong, and in this case and the prediction will also go wrong.

So, if I consider that, if I ignore this multicollinearity what can happen is that my prediction model will show something different and actual scenario may be something else like that. So, in this case, I need to rectify this multicollinearity, there are different ways of rectifying the multicollinearity problem.

So, one of the one of the approach that takes care of this may be this what we are using as what we are using there are two methods over here. So, one of the method that is stepwise regression we have adopted like that and that may eliminate this multicollinearity problem that we are having. So, an another method is known as best subset regression based on which we can select variables which will go in and which will go out like that.

So, first is best substitute that we can talk about is stepwise regression like that. So, what we will do is that we will go to regression. So, when we have fitted this model that this variation inflation factor that you see over here. What; if I copy this one as image and we can paste it in excel, and let us try to see enlarge the image and try to see what is the results over here.

(Refer Slide Time: 12:31)



Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	62.4	70.1	0.89	0.399	
x1_1	1.551	0.745	2.08	0.071	38.50
x2_1	0.510	0.724	0.70	0.501	254.42
x3_1	0.102	0.755	0.14	0.896	46.87
x4_1	-0.144	0.709	-0.20	0.844	282.51

So, variation inflation factor for each of these variables is indicative over here in the VIF, what we are seeing and we can just go to a simple and we can paste this one. And what we see over here is basically 38.5, X1 is having a variation inflation factor of 38.5, this is 254, this is 46, there is high amount of correlation that exists between the x variables over here. Which one is highly correlated which one we can see by the correlation matrix plot, and we will be able to know which is related with which one.

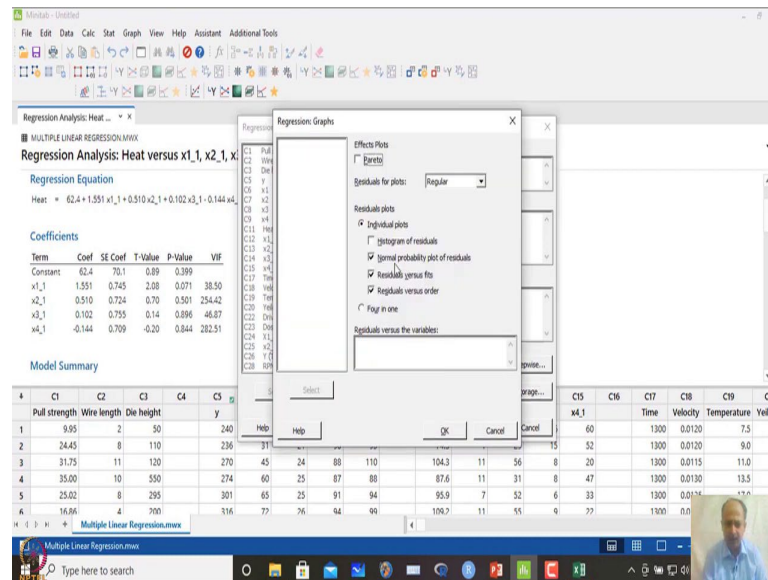
So, whenever this relationship strong relationship exists this variation inflation factor will be more than 5, will be more than 5 or 10 like that. So, generally statistician follows some rule, or thumb rules like that. If it is more than 5 we will take action and we want to eliminate multicollinearity problem in the regression equation.

So, that my prediction model becomes more accurate like that. So, anything more than 5 is we may consider a significant, we can take actions over there by using different methods and addressing the regression equations like that, developing the best regression equation like that.

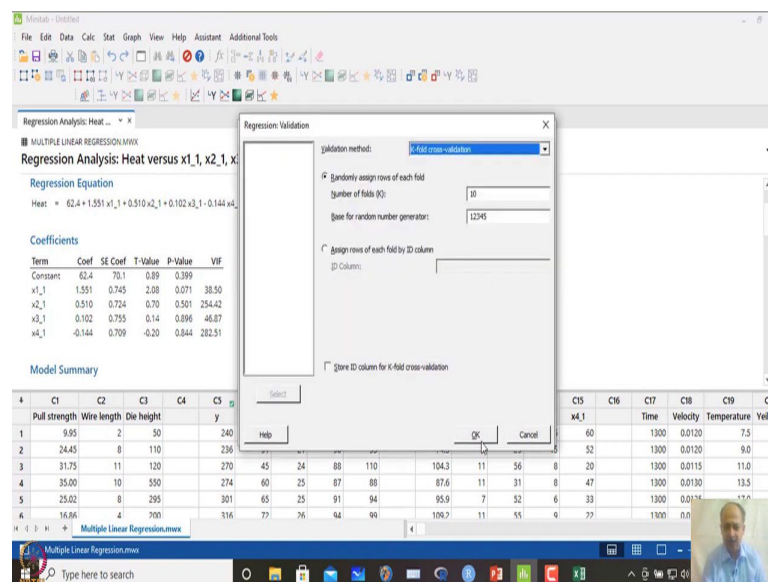
So, here problem is that it is more than 10 or criteria 5 whatever you can select like that. And generally recommended is 10, but some statistician also suggest anything more than 5 is also a concern for me, so we should yeah we should try to remove that multicollinearity problem and then develop the regression equation like that, ok. So, this is a problem variation inflation factor.

And in this case the options that we are having is that we will go for a stepwise equation. So, what we will do regression fit regression model over here. So, this is X1 to X4. So, in this case I will use stepwise regression and let us try to see how it works in this case.

(Refer Slide Time: 14:15)



(Refer Slide Time: 14:21)



And then I will go for let me see storage of this residual is already there and I will click and graph. What we can do is that? We can see residual process with normal probability plot after doing stepwise regression, validation tenfold cross-validation we are doing over here.

(Refer Slide Time: 14:24)

Regression Analysis: Heat versus x1_1, x2_1, x3_1, x4_1

Method
Cross-validation 10-fold

Stepwise Selection of Terms
a to enter = 0.15, a to remove = 0.15

Regression Equation
Heat = 52.58 + 1.468 x1_1 + 0.6623 x2_1

Coefficients

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1_1	x2_1	x3_1	x4_1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120		7.5
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120		9.0
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115		11.0
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130		13.5
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0130		13.0
6	16.86	4	700		316	77	26	94	99		106.7	11	55	9	77		1300	0.0130		13.0

(Refer Slide Time: 14:27)

Regression Analysis: Heat versus x1_1, x2_1, x3_1, x4_1

Stepwise Selection of Terms
a to enter = 0.15, a to remove = 0.15

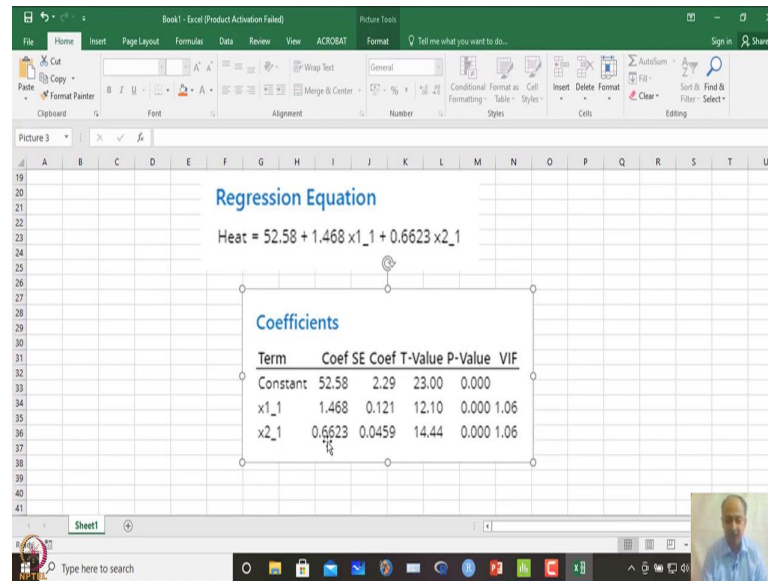
Regression Equation
Heat = 52.58 + 1.468 x1_1 + 0.6623 x2_1

Coefficients

Term	Coeff	SE Coef	T-Value	P-Value	VIF
Constant	52.58	0.29	23.00	0.000	1.06
x1_1	1.468	0.121	12.10	0.000	1.06
x2_1	0.6623	0.0459	14.44	0.000	1.06

And I click ok. What we observe is that only X1 and X2 is retained over here. So, you see the equation after doing stepwise regression what is coming out on this model.

(Refer Slide Time: 14:38)

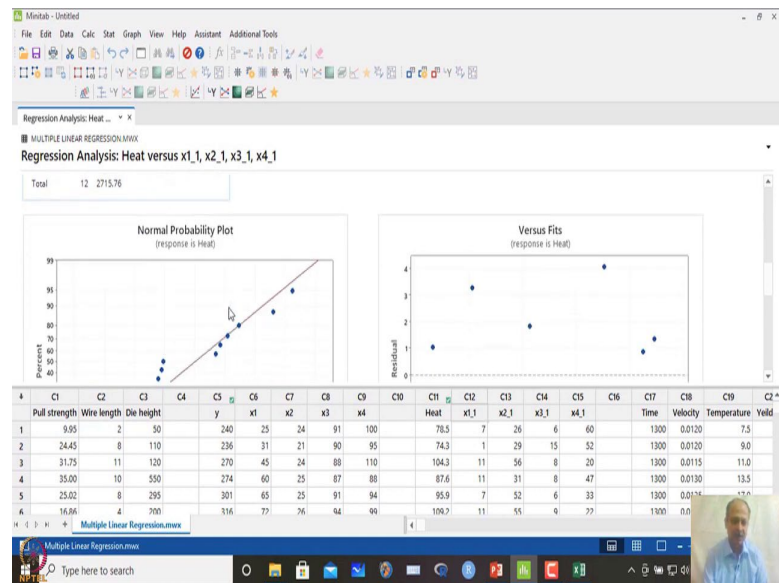


And we are getting a regression equation like this and heat is equal to 52.58 into 1.468. This is a coefficient plus coefficient what we have for X1 and X2. And the results also indicate that, now variation inflation factor if you look at this column what happened is that copy as picture and if we have taken this as the final equation, then the variation inflation vector is near to 1.

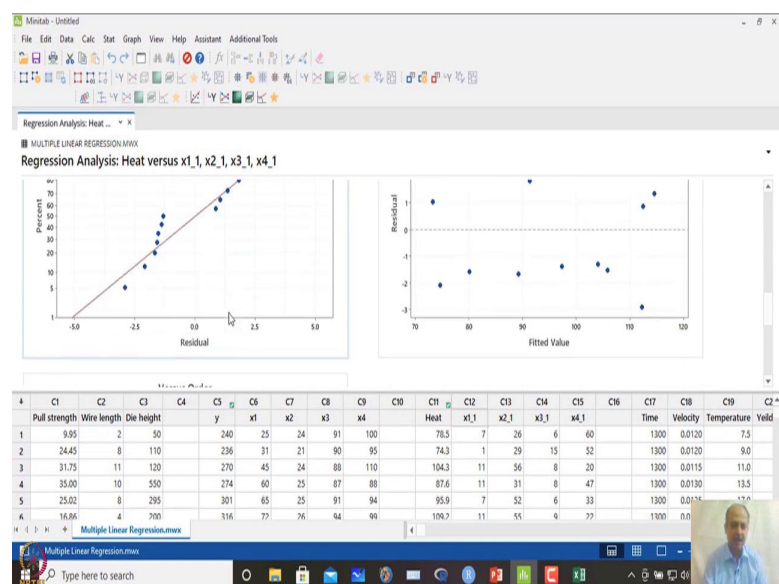
What you can see if we if I enhance this one that this is near to 1. When it is near to 1, that means, it is quite perfect and no multicollinearity problem does not exist now, and X1 and X2 are independent over here. We can assume independence in between the variables over here.

So, we have replaced this is X3 and X4 basically. We have just removed X3 and X4 over here. And the R square value and tenfold cross-validation more or less they are close to each other.

(Refer Slide Time: 15:31)

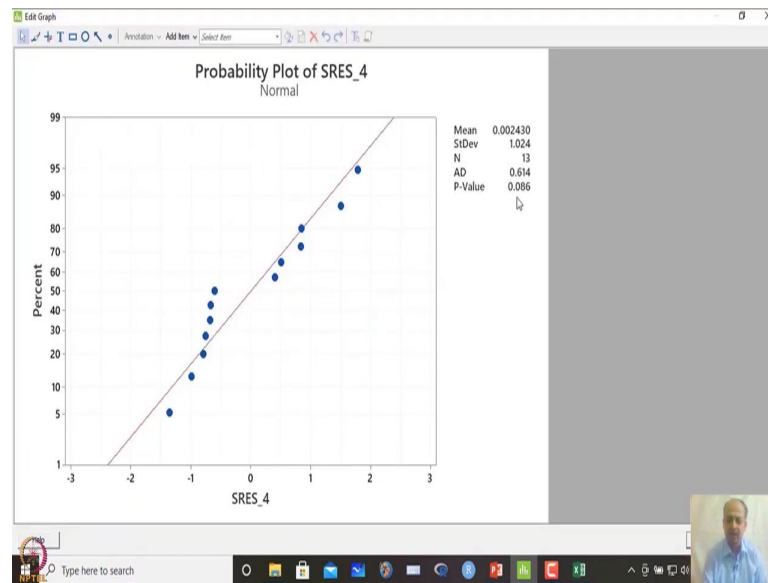


(Refer Slide Time: 15:32)



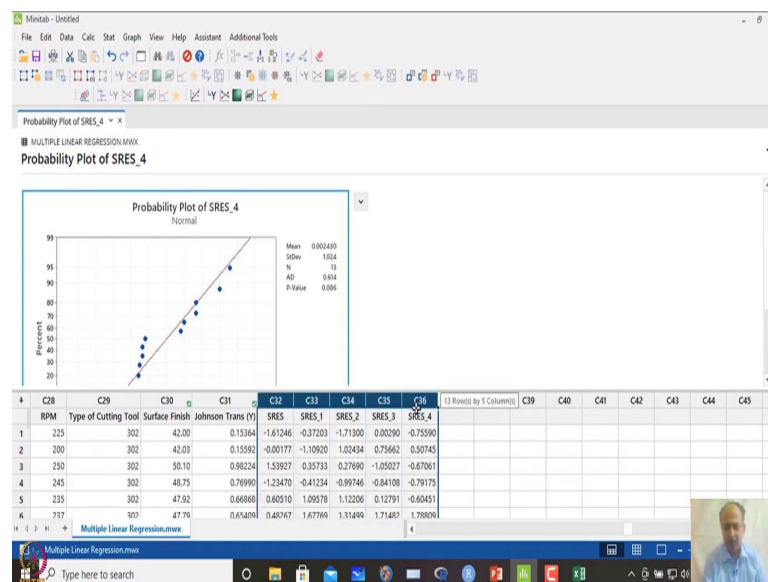
And ANOVA analysis also shows they are significant. And we can see that residual normal probability plot and we can also verify whether the final residual is normal or not, it is falling normal or not.

(Refer Slide Time: 15:41)



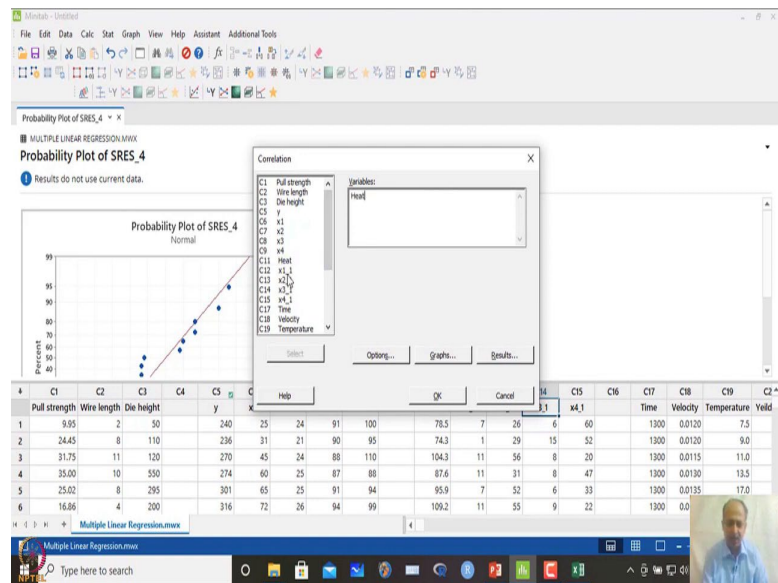
So, what we observe over here is that, ok, this assumption is also validated. So, P is more than 0.05. So, this is there is no problem in the error or residuals like that. So, we can just remove for our benefit, so that later on we have only required information over here.

(Refer Slide Time: 15:54)

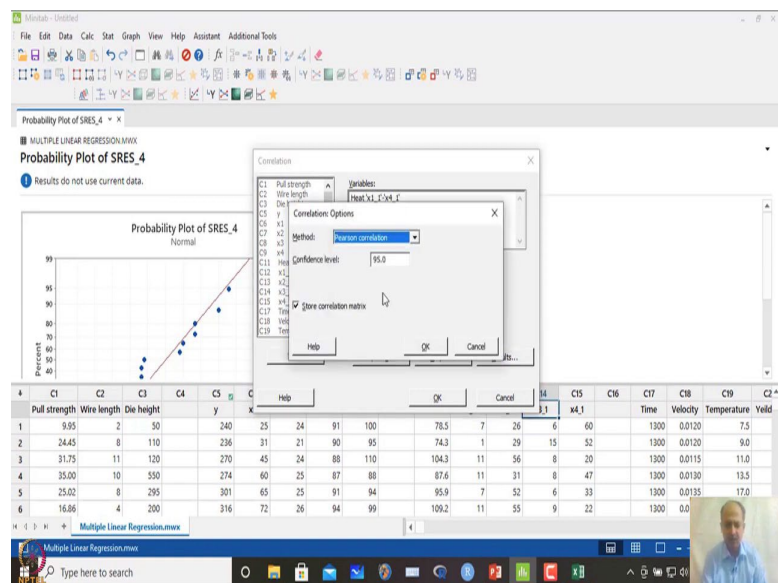


So, what we can do is that. So, this is verified over here. So, two variable goes in and two variable goes out over here. So, that means, this stepwise regression has taken care of this multicollinearity problem over here, somewhat we are fortunate that this is taken care of over here.

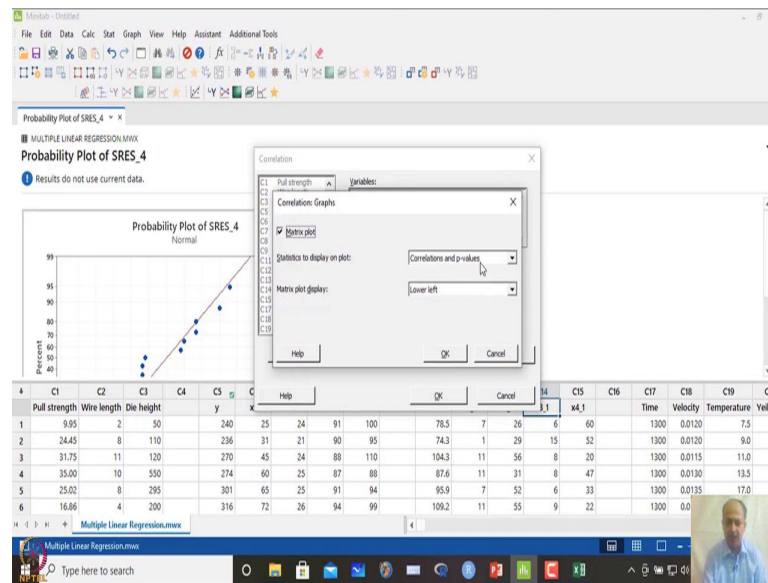
(Refer Slide Time: 16:32)



(Refer Slide Time: 16:37)



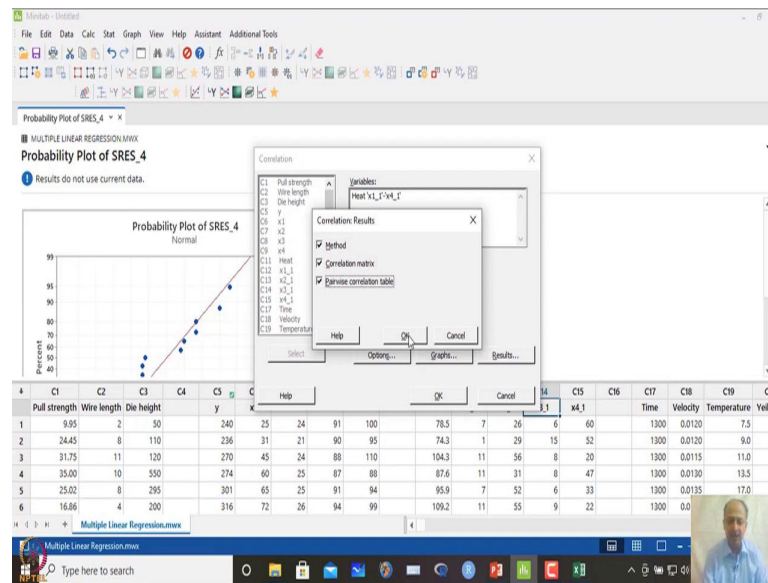
(Refer Slide Time: 16:43)



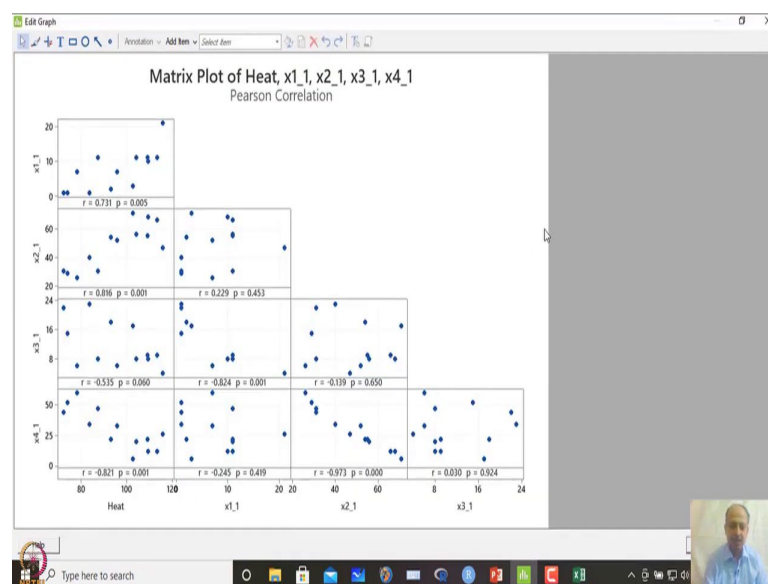
So, why; and let us try to see multicollinearity issues over here. So, what we can do is that basic stat over here and we go to correlation over here. So, correlations between what we can do is that we can see the correlation between heat and other variables over here from X1 to X4, and we can select this one I want to see the correlation matrix.

So, I go to options and I say Pearson correlation I want to say, I want to identify. And then in graphs what we can do is that I in the plot I want correlation with P values like that because I am interested in P values which is correlated with which one and which is significant like that. And we can see a pairwise correlation also.

(Refer Slide Time: 16:53)



(Refer Slide Time: 16:56)



And when I click ok what will happen you will get some graphs like this. So, this graph will indicate how what is the correlation between y variable and X1, X2 and inter relationship between X1 and X2, how is the correlation?

So, if I see the first column over here what you see heat is having showing a P value of 0.01 with X4 variables. So, heat is highly correlated with X4, heat is highly correlated with X3, not so much it is more than 0.05. So, we can say that this may not be significant

so much. But X_2 i C p value is less than 0.05 X_1 , X_1 , X_2 and X_4 there are there is highly high amount of correlation that is existing.

Now, X_1 if you see, X_1 is highly if I see within the variables. So, X_1 how it is related with X_4 , it does not have any correlation X_4 . But with X_3 it is having a high level of correlation over here. So, X_1 is related with X_3 ; X_1 and X_3 are highly correlated over here. So, in this case, this is the observation.

Then, similarly X_2 you see perfect relationship exist between X_2 and X_4 . So, X_2 and X_4 are more or less perfect and the r coefficient is negative coefficient is 0.973. So, negative correlation exists between this variable X_1 , X_2 and X_4 . So that means, these two variables are highly correlated. Similarly, X_1 and X_3 is highly correlated.

So, whenever high correlation exists and I want to do regression in that case what is required is that one of the variable has to go, one of the variable has to go out of this X_1 and X_3 . We can think of an X_2 and X_4 basically has to go, one of the variables has to go. And the stepwise regression has correctly identified two variables instead of 4 and it has identified X_1 and X_2 .

It has retained those two variables and remove the other two variables because there is a multicollinearity problem. So, some part of multicollinearity problem can be addressed by when I use stepwise regression. But that can always be verified by seeing the variation inflation factor and seeing the model adequacy and other checks. And finally, we select the models over here.

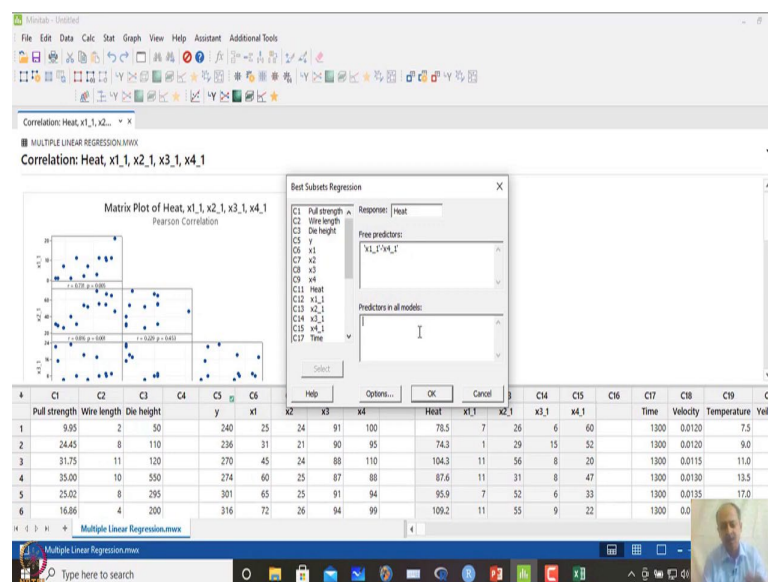
So, this correlation matrix helps you to understand that which variable is highly correlated to which one which variable is removed like that which variable is added by. So, this is one way of selecting the variables which is known as stepwise regression when there is a multicollinearity problem that is existing over here, ok.

So, this and there is another option which can be explored over here which is known as best subset regression which is another option. This stepwise regression what the limitation of this approach is that it will select the final variables X_1 and X_2 which is selected like that, but it will it has dropped X_3 and X_4 .

But scenario can be that I want to explore what happens if I include X3 instead of X1; what happens if I include X4 instead of X2 because those variables are easy to control maybe because I want a regression equation where variables can be easily controlled like that maybe X1 and X2 is too difficult, but regression equation by significance and best subset methodology we are getting X1 and X2.

But I want to see the complexity if I use different combination like that. If I use X3 X4 combination what is happening and like that; so what will happen. So, there is option which is known as best subset regression. So, if you go to regression, regression in MINITAB you will see an option of best subset over here.

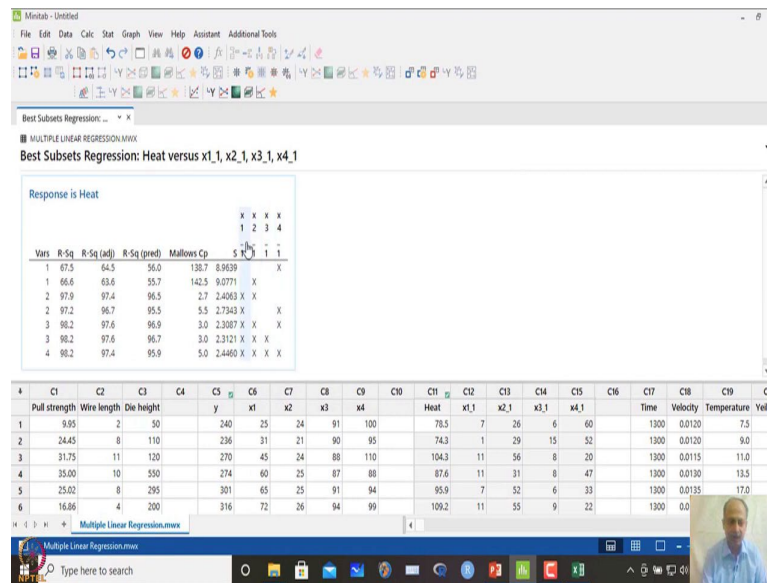
(Refer Slide Time: 20:18)



So, if you click best subset regression for this then you identify which is the variable. So, I give heat as the response over here and then X1 to X4, these are the variables I want to include in the model which is free predicted over here. So, if you want some predicted to be always there in the model, so predicted in all models. So, in that case you can just write X1 to be there.

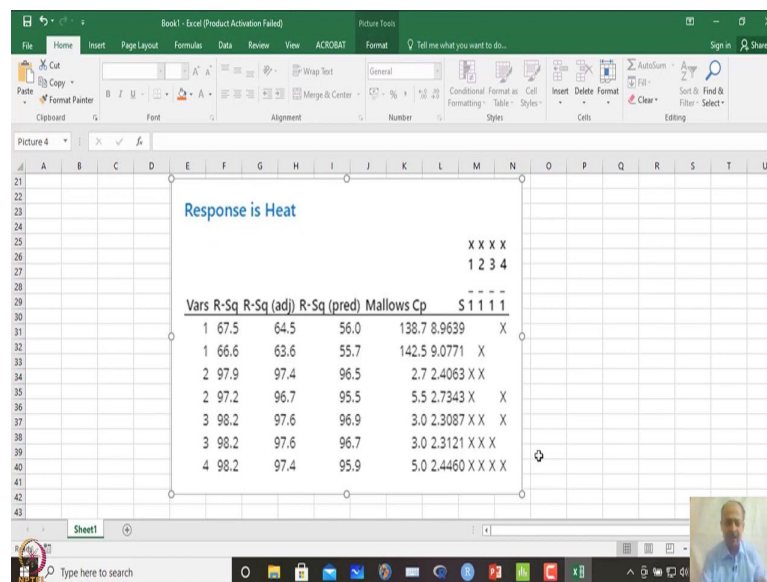
So, I am not identifying that one. So, I said every variable is free. So, show me all combinations and best combinations like that.

(Refer Slide Time: 20:48)



So, I click ok over here, then you will get a information over here which I am copying and I will just paste this information over here. So, that it is easier to see also.

(Refer Slide Time: 21:03)



So, in this case, what you observe is that there is an indicator which is known as Mallows Cp over here which is known as Mallows Cp over here. So, this Mallows Cp is generally used to select the best model and combination. So, with one variable combination; so, over here what you see is that there are X1, X2, X3, X4 variables over here.

And there are with one variable that was selected the best model with one variable that is given over here. So, first two is giving you the best model. So, if I include X4 that is giving me a high value of R square adjusted over here and high value of R square also over here and this is the best model with one variable that is selected and that is the summary of that is given over here.

Similarly, with second best model that is with R square value 66.6 and I am getting a variable that should be that should be retained only X2. So, if you model with X4 what is happening; when model with X2 what is happening; best two models are given.

So, with single variable we have 4, so 4 models can be developed like that, 4 models can be developed like that. So, if you have k number of variables two to the power k is the total combinations of the permutation combination we can think of, all possible models, all possible models that can happen like that.

So, MINITAB only reports that best model with 1 variable, best model with 2 variable, best model with 3 variable. And finally, with all variables like that. So, we will ignore the last one because we want to reduce the number of variables and we will go by the lower combination which is giving you better options because I want to reduce the number of variables. Because I want to control less number of variables like that, and also I want to take care of multicollinearity that is an issue, ok.

So, in this case what we do is that there are 3 values that we see R square adjusted to make a compromise over here, R square adjusted, R square predicted and Mallow Cp. First we go by Mallow Cp in the indicator over here which indicates the variance is less. And in this case the error that is that is the sum of square error that we are committing over here that is that we are getting over here is much less as compared. So, how do you how do we compare that one?

There is an indicator that is known as Mallow Cp which says that the thumb rule over here is that the Mallow Cp value should be less than the number of variables considered for modeling plus 1. So, that is if you are going for let us say for first one over here, so this is number of variable considered is 1, X4 is considered over here in the model.

So, X4 plus 1 that is that is one variable that is 1 plus 1 is 2, and the Mallow Cp should be less than the 2 less than the value 2 over here and that is not the case over here 138.7

is the Mallow Cp indicator that we are getting over here. So, this criteria until and unless this criteria is fulfilled. So, this is not the best model that we should select.

Secondly, second one also you see that 142.5 which is very very higher than 2, that also can be eliminated over here. But in the third case what you see is that two variable model when X1 and X2 is considered Mallow Cp value is approximately 2.7. So, in these case this calculation of Mallow Cp any books will give you what is the calculation values of Mallow Cp considering, sum of square errors over here.

So, the formulation is given. So, I am not mentioning the formulation over here. That you can see in any standard textbook. So, what I am recommending is that this value you see two number of variables, so total number of 2.7 is less than 3.

So, this can be a possible combination. And also stepwise regression has also shown that this is the combination X1 and X2 is the best combination and that is the best model that stepwise regression has identified like that. Here also the suggestion is Mallow Cp is less than and very close to the number of variables plus 1. So, this value should be very close to number of variables plus 1. So, 2.7 is close to 3, so that means, this is one of the competitor models over here, ok.

And this is 5.5 it is more than, so this also goes, and this 3 is also less than 4 over here, so this can be one of the possibilities and this is another one possibilities over here. Now, you have to check that whether this model is with two variables is good with 3 variables what is happening with other 3 variables what is happening like that.

Because of correlation that exists between these two you will find that variation nucleation factor will be high, whenever I consider X1 and X3 together or X2 and X4 together like that there will be problems like that.

So, over here the closest model; that means based on Mallow Cp criteria also we are seeing that X1 and X2 is the best one. Mallow Cp is 2.7 R square predicted is 96 that is quite good enough, R square adjusted is 97.4. So, and this seems to be the closest one and we should select this one.

So, based on Mallow Cp criteria and based on our stepwise regression we are converging to the same models which can be suggested over here, which is the best model over here.

So, X1, X2 variables regressed with y that is the best model over here. So, if you have if we have considered any other combination of that. So maybe we will not get the best models or maybe the assumptions will be violated, assumptions of normality and other assumptions that is there that can be violated head to schedule so, all these things.

So, whenever I have selected the best models there is always a requirement for checking the model adequacies. So, all the error terms and the assumptions of the errors are to be verified like that. So, in this case, what we see; and we see that X1 and X2 is the best selection over here, ok. We can check what is happening what is happening if I select X1, X2, and X4 over here? But X2 is highly related with X4. So, if you have selected that one variation inflation factor that problem will come.

(Refer Slide Time: 26:51)

The screenshot shows the Minitab software interface. The main window displays the 'Best Subsets Regression: Heat versus x1, x2' results. The 'Response is Heat' is indicated. The table below shows the results for the best subsets regression.

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp
1	67.5	64.5	55.0	138.7
2	97.9	97.4	96.5	2.7
3	98.2	97.6	96.9	3.0
4	98.2	97.4	95.9	5.0

The 'Regression' dialog box is open, showing the 'Responses' list with 'Heat' selected. The 'Continuous predictors' list contains 'x1, x2, x4'. The 'Categorical predictors' list is empty. The 'Model...' button is highlighted.

The bottom of the screenshot shows the 'Multiple Linear Regression.mnx' file open, displaying a table of data with columns C1 (Pull strength), C2 (Wire length), C3 (Die height), C4 (Yield), and C5 (Yield).

(Refer Slide Time: 26:58)

Regression Stepwise

Method: None

Potential terms:

- X1
- X2
- X3
- X4

Alpha to enter: 0.15

Alpha to remove: 0.15

Best Subsets Regression: Heat versus x1, x2

Response is Heat

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp
1	67.5	64.5	56.0	138.7
2	97.9	97.4	96.5	2.7
3	98.2	97.6	96.7	3.0
4	98.2	97.4	95.9	5.0

	C1	C2	C3	C4	C5
Pull strength	Wire length	Die height			y
1	9.95	2	50		240
2	24.45	8	110		236
3	31.75	11	120		270
4	35.00	10	550		274
5	25.02	8	295		301
6	16.86	4	200		316

	C15	C16	C17	C18	C19	C20
x4,1	Time	Velocity	Temperature	Yield		
60	1300	0.0120	7.5			
52	1300	0.0120	9.0			
20	1300	0.0115	11.0			
47	1300	0.0130	13.5			
33	1300	0.0135	17.0			
22	1300	0.0				

So, if I assume this one. So, in this case regression, if I assume X1, X2 and X4. So, in this case, if I remove this one X1, X2 and X4 and X2 and X4. We have seen highly correlated and I remove stepwise regression over here.

(Refer Slide Time: 27:01)

Regression Analysis: Heat versus x1, x2, x4,1

Method

Cross-validation 10-fold

Regression Equation

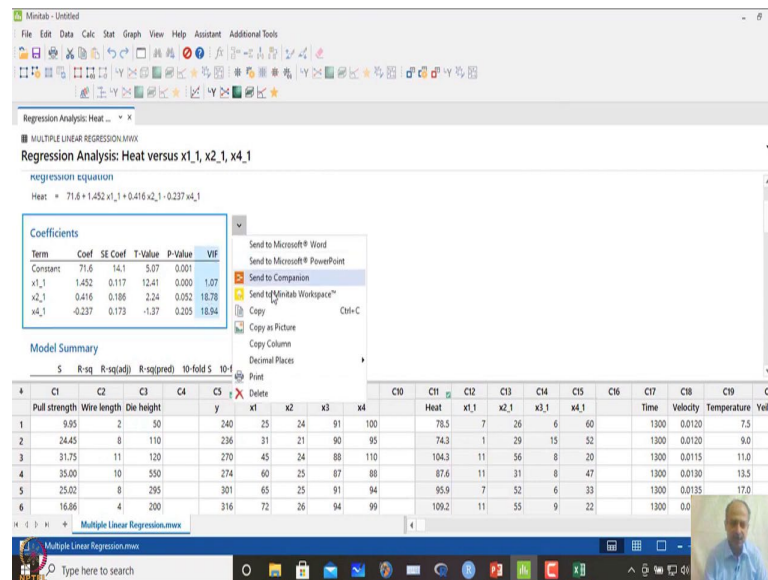
Heat = 71.6 + 1.452 x1,1 + 0.416 x2,1 - 0.237 x4,1

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	71.6	14.1	5.07	0.001	
x1,1	1.452	0.117	12.41	0.000	1.07
x2,1	0.416	0.186	2.24	0.052	18.78

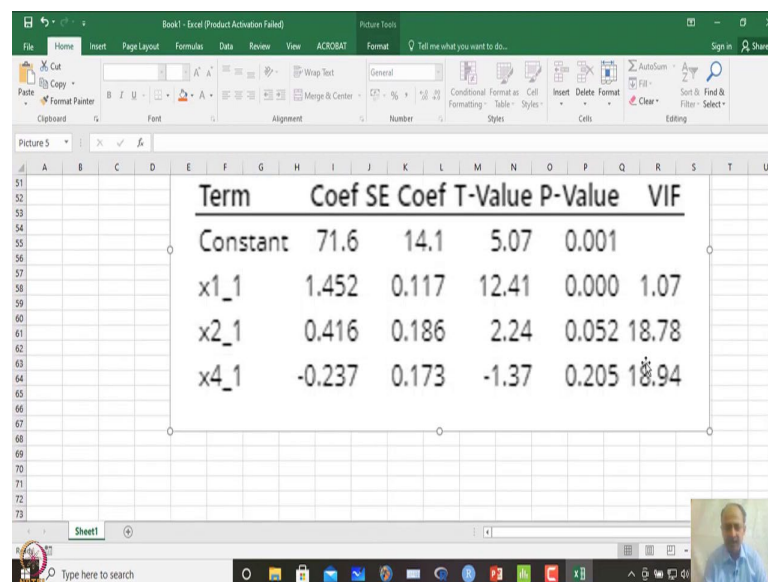
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
Pull strength	Wire length	Die height			y	x1	x2	x3	x4		Heat	x1,1	x2,1	x3,1	x4,1		Time	Velocity	Temperature	Yield
1	9.95	2	50		240	25	24	91	100		78.5	7	26	6	60		1300	0.0120	7.5	
2	24.45	8	110		236	31	21	90	95		74.3	1	29	15	52		1300	0.0120	9.0	
3	31.75	11	120		270	45	24	88	110		104.3	11	56	8	20		1300	0.0115	11.0	
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130	13.5	
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0135	17.0	
6	16.86	4	200		316	72	26	94	99		109.2	11	55	9	22		1300	0.0		

(Refer Slide Time: 27:04)



And then, I do the calculation over here. What is being observed you see that the variation inflation factor that you observe over here is very high; variation inflation factor is very high over here.

(Refer Slide Time: 27:10)



If I paste this one over here, and you will find that variation inflation factor is 18.78, X2 and X4, there is high correlation that exists between X2 and X4 that was prominent in correlation coefficient also. So, this whenever it is high this type of regression equation cannot be used that is the overall suggestion that we have, ok.

So, and in this case what happens is that y. So, let us take this example to finish off with this y equals 2. And we will continue with another examples and another situation in multi multiple regression when the error assumption fails in that case what we can do like that. And then we go into the design of experiment part of that.

So, why I am explaining this because when we develop a design of experiments regression equation we will we should be concerned about the variable interrelations between the variables, and how to select the best models out of many variables like that, how to eliminate variables like that, ok.

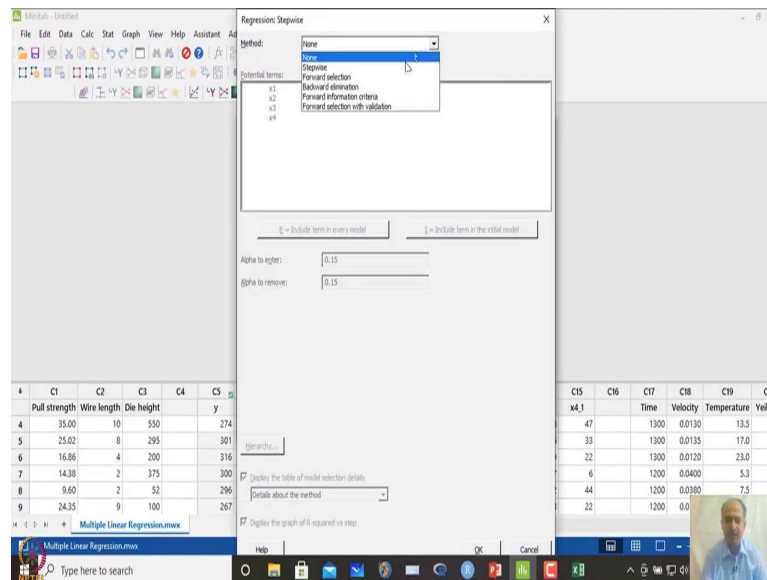
If this is the scenario y and X1 and X4, so we can eliminate this one and we go by this regression analysis, regression over here, fit regression models.

(Refer Slide Time: 28:20)

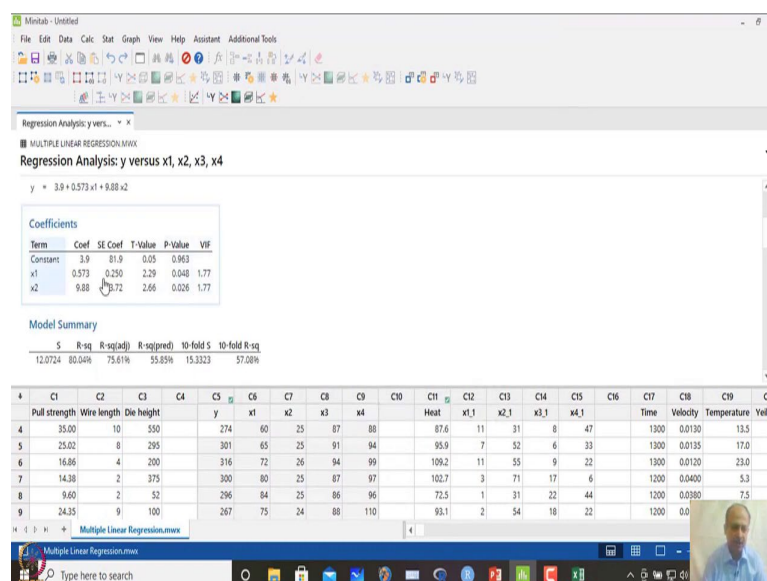
The screenshot shows the Minitab software interface. A 'Regression' dialog box is open, allowing the user to select variables for a regression model. The 'Responses' field is set to 'y'. Under 'Continuous predictors', 'x1' and 'x4' are selected. The 'Categorical predictors' field is currently empty. Below the dialog box, a portion of a data table is visible, showing columns C1 through C24. The first row of data includes values for 'Pull strength', 'Wire length', 'Die height', and 'y'.

#	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
4	Pull strength	Wire length	Die height		y																			
5	35.00	10	550		274																			
6	25.02	8	295		301																			
7	16.86	4	200		316	72	26	94	99															
8	14.38	2	375		300	80	25	87	97															
9	9.60	2	52		296	84	25	86	96															
10	24.35	9	100		267	75	24	88	110															

(Refer Slide Time: 28:30)



(Refer Slide Time: 28:36)

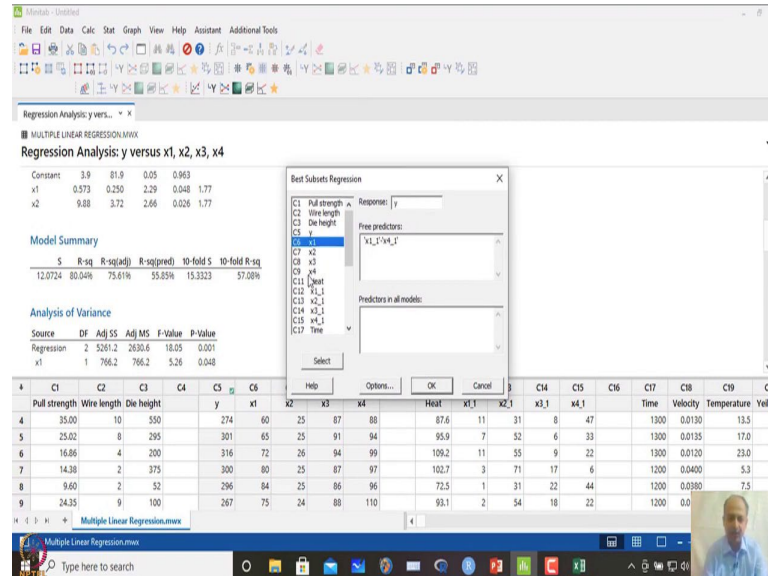


So, in this case, what happens is that I selected y variables over here and then I select X1 to X4 and select this sorry this has to be coming over here continuous predictor X1 to X4 and I select this one and stepwise regression what we have done is that we use stepwise regression over here and use this one.

So, suggested model is X1 and X2. These are the variables. And variation inflation factor is less, so this can be the best model. Only thing is that R square adjusted is around 74,

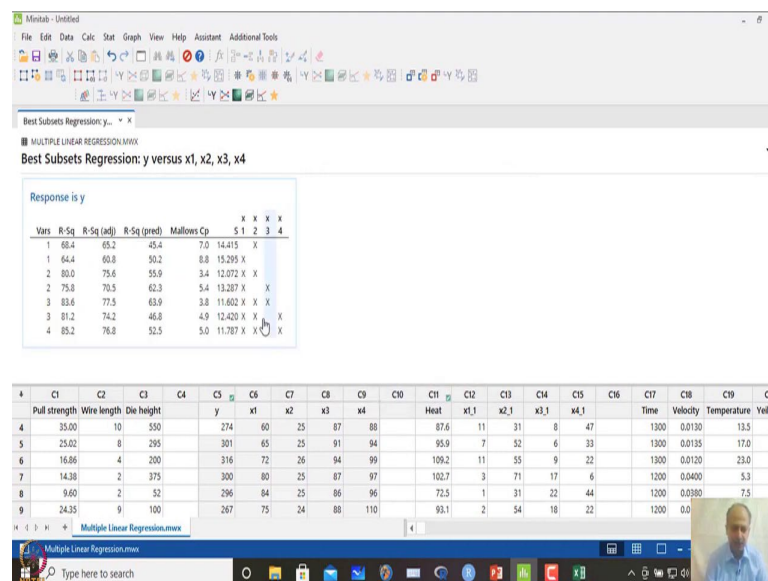
75.61 and this is 57. So, there is some gap that we are observing over here 75 and 57 over here. So, whether we can improve this tenfold cross-validation over here.

(Refer Slide Time: 29:01)



So, again what we can do is that we can see this regression by best subset values. And the best subset regression we can do with y and X1 to X4 and try to see what the model recommendation like that.

(Refer Slide Time: 29:10)



So, in this case what you see is that X1 and X2 which is best subset is giving me a value of 3.4. So, this is about 3 over here this is about. So, this is more than 3 basically. So,

number of variables plus 1 and this is more than 3 over here. So, and this is 3.8 which is very close you see 1, 2, 3, 4.

So, X_1 , X_2 and X_3 variables if I consider and that is coming out to be very close. So, Mallow C_p , based on Mallow C_p index what we are seeing is that if I consider X_1 and X_2 and X_3 variables over here that is giving me a Mallow C_p which is approximately 3.8, which is very close to 4 and in that case 3 variables can be considered.

(Refer Slide Time: 29:52)

The screenshot shows the Minitab software interface. The main window displays the 'Best Subsets Regression' results for the response variable 'y' versus predictors 'x1', 'x2', and 'x3'. The table shows the following data:

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp
1	68.4	65.2	45.4	7.0
2	80.0	75.6	55.9	3.4
3	83.6	77.5	63.9	3.8
4	85.2	78.8	62.5	5.0

The 'Regression' dialog box is open, showing the response variable 'y' and the continuous predictors 'x1', 'x2', and 'x3'. The 'Model' button is highlighted.

So, if I go over to 3 variables over here fit 3 variables; so, let us reduce this one, let us incorporate X_1 , X_2 and X_3 . Stepwise says X_1 and X_2 only. So, we will remove stepwise over here and try to see what the model gives.

(Refer Slide Time: 30:03)

Regression Analysis: y versus x1, x2, x3

Method
Cross-validation 10-fold

Regression Equation
 $y = -162 + 0.749x_1 + 7.69x_2 + 2.34x_3$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-162	148	-1.09	0.306	
x1	0.749	0.275	2.73	0.026	2.32
x2	7.69	3.94	1.95	0.087	2.16
x3	2.34	1.77	1.32	0.223	1.32

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	10-fold S	10-fold R-sq
11.6017	83.62%	77.47%	63.91%	14.2406	62.98%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Constant	1	14240.6	14240.6	1.09	0.306
x1	1	14240.6	14240.6	2.73	0.026
x2	1	14240.6	14240.6	1.95	0.087
x3	1	14240.6	14240.6	1.32	0.223
Error	10	14240.6	1424.06		
Total	21	14240.6			

(Refer Slide Time: 30:06)

Regression Analysis: y versus x1, x2, x3

Method
Cross-validation 10-fold

Regression Equation
 $y = -162 + 0.749x_1 + 7.69x_2 + 2.34x_3$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-162	148	-1.09	0.306	
x1	0.749	0.275	2.73	0.026	2.32
x2	7.69	3.94	1.95	0.087	2.16
x3	2.34	1.77	1.32	0.223	1.32

Model Summary

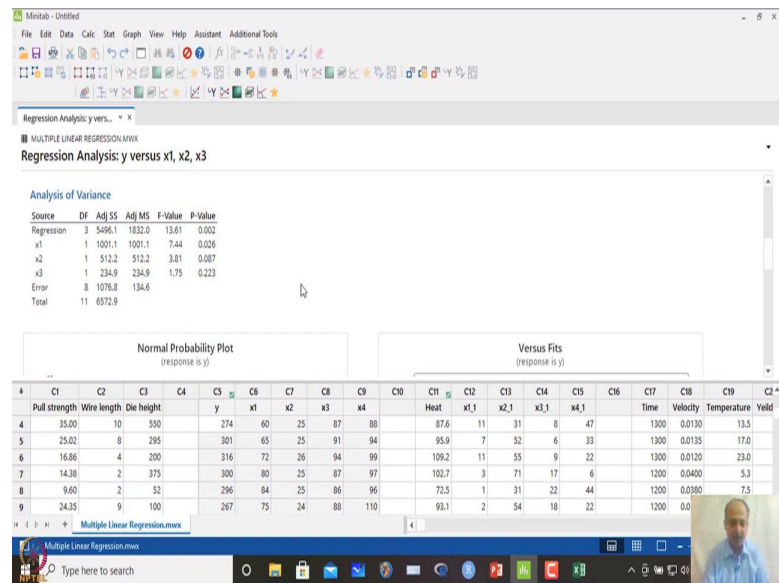
S	R-sq	R-sq(adj)	R-sq(pred)	10-fold S	10-fold R-sq
11.6017	83.62%	77.47%	63.91%	14.2406	62.98%

Analysis of Variance

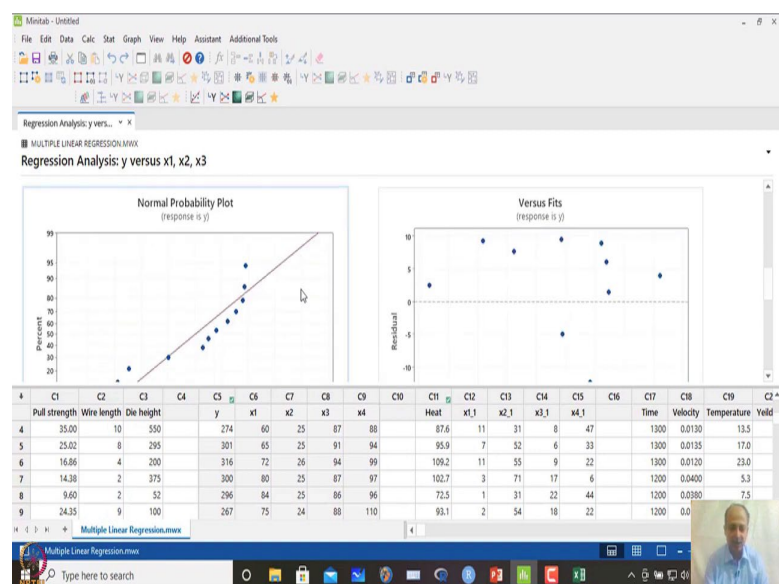
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Constant	1	14240.6	14240.6	1.09	0.306
x1	1	14240.6	14240.6	2.73	0.026
x2	1	14240.6	14240.6	1.95	0.087
x3	1	14240.6	14240.6	1.32	0.223
Error	10	14240.6	1424.06		
Total	21	14240.6			

So, if you click ok over here what happens is that it gives a 3 model 3 variable models over here. So, over here what you see is that only X1 is coming prominent and others two are not coming prominent because the P value is more than 0.05 over here. Although the variation inflation factor is not significant over here, but R square predicted R square adjusted value somewhat improved and the tenfold cross-validation is also somewhat improved over here 62.98, ok.

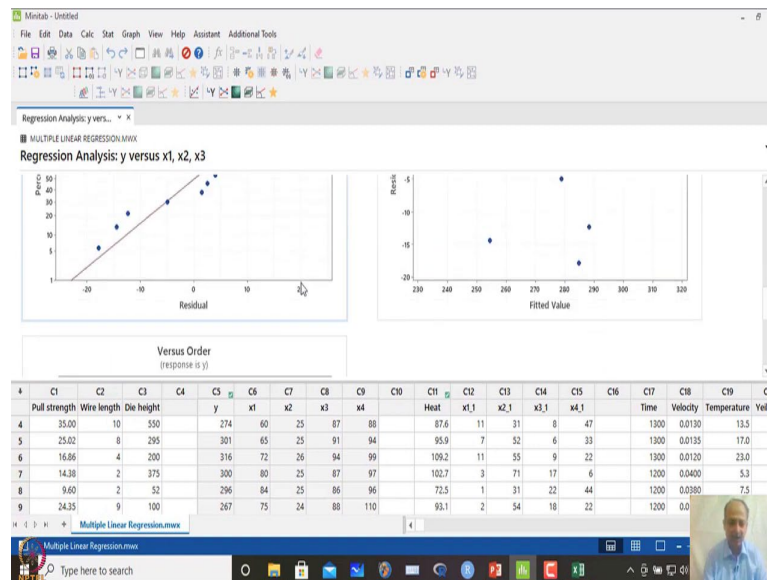
(Refer Slide Time: 30:29)



(Refer Slide Time: 30:31)

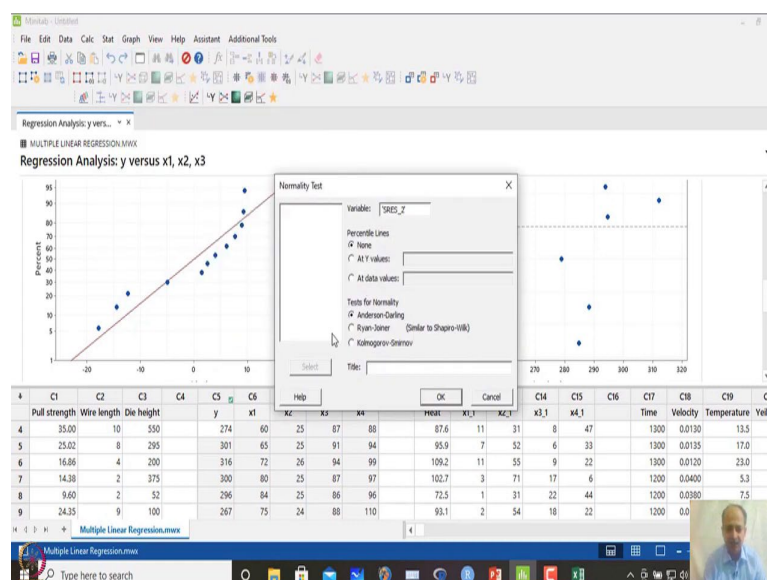


(Refer Slide Time: 30:31)

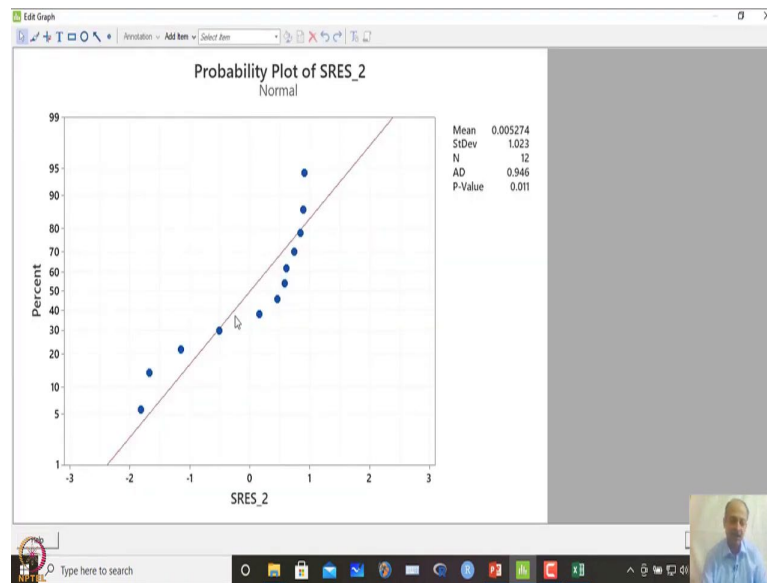


But the residuals that I have saved over here. So, this is the residual plot that you see. So, when I have used 3 variable models over here what happens is that; if I go to basic stat normality test what will happen is that. So if I go to the last variables and try to test this one what happens is that you see that there is a violation in the error distribution over here.

(Refer Slide Time: 30:42)

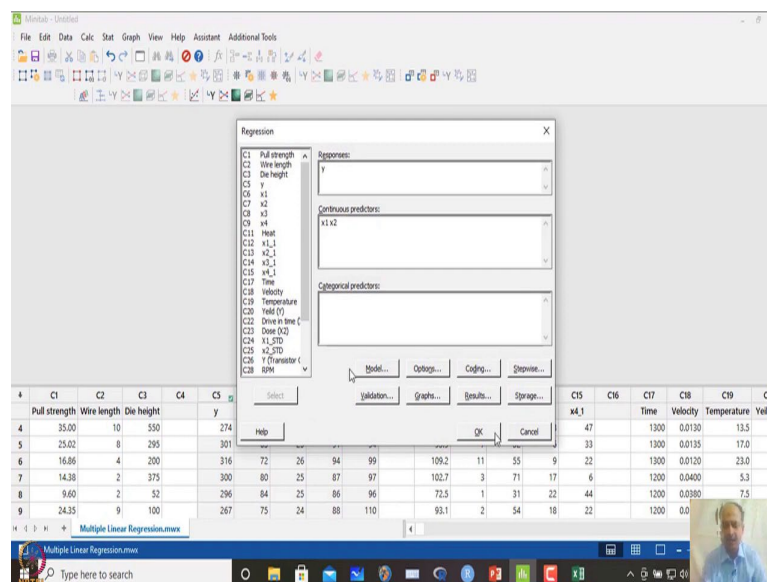


(Refer Slide Time: 30:45)



So, whenever I have added this one. So, if I have restricted this to 2 variable models like that. So I go to basic regression analysis regression fit regression model instead of X3, I go to X1 and X2 only which is suggested by stepwise regression.

(Refer Slide Time: 30:58)



(Refer Slide Time: 31:02)

Regression Analysis: y versus x1, x2

Method
Cross-validation 10-fold

Regression Equation
 $y = 3.9 + 0.573 x_1 + 9.88 x_2$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.9	81.9	0.05	0.963	
x1	0.573	0.250	2.29	0.048	1.77
x2	9.88	3.72	2.66	0.026	1.77

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
	Pull strength	Wire length	Die height		y	x1	x2	x3	x4		Heat	x1_1	x2_1	x3_1	x4_1		Time	Velocity	Temperature	Yield
4	35.00	10	550		274	60	25	87	88		87.6	11	31	8	47		1300	0.0130	13.5	
5	25.02	8	295		301	65	25	91	94		95.9	7	52	6	33		1300	0.0135	17.0	
6	16.86	4	200		316	72	26	94	99		109.2	11	55	9	22		1300	0.0120	23.0	
7	14.38	2	375		300	80	25	87	97		102.7	3	71	17	6		1200	0.0400	5.3	
8	9.60	2	52		296	84	25	86	96		72.5	1	31	22	44		1200	0.0380	7.5	
9	24.35	9	100		267	75	24	88	110		93.1	2	54	18	22		1200	0.0		

And I save the residual over here and I go the I do the normality test over here with the residuals residual 3 which is saved over here. And I do this and what I see is that the residuals are perfectly following normal distributions like that 0.253 like that.

(Refer Slide Time: 31:07)

Regression Analysis: y versus x1, x2

Method
Cross-validation 10-fold

Regression Equation
 $y = 3.9 + 0.573 x_1 + 9.88 x_2$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.9	81.9	0.05	0.963	
x1	0.573	0.250	2.29	0.048	1.77
x2	9.88	3.72	2.66	0.026	1.77

Normality Test

Variables: SALES_T

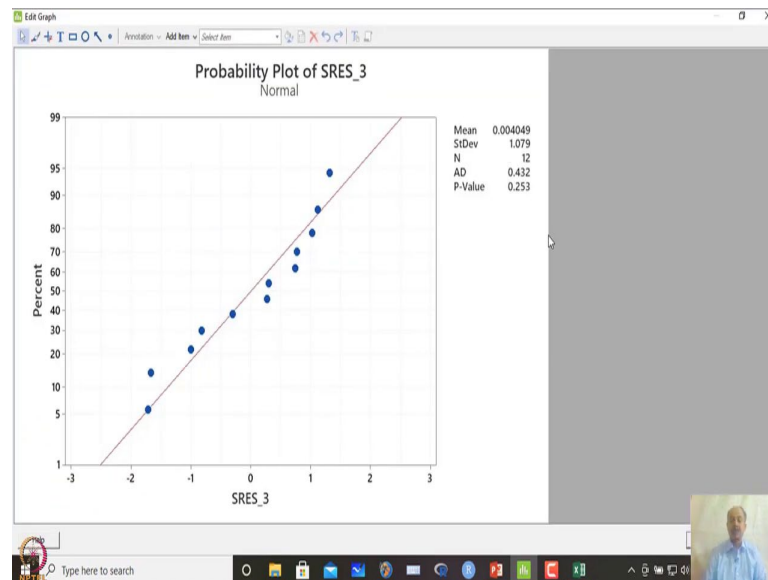
Percentile Lines
☐ None
☐ ALLY values:
☐ At data values:
☐

Tests for Normality
☒ Anderson-Darling
☐ Ryan-Jensen (Similar to Shapiro-Wilk)
☐ Kolmogorov-Smirnov

Title:

OK Cancel

(Refer Slide Time: 31:10)



So, we will always go by suggestions that is what statistician has suggested. So, we go by stepwise regression. We do not add unnecessary variables which are not non-significant terms, but whenever I am removing a non-significant terms please remember that we are losing some amount of information.

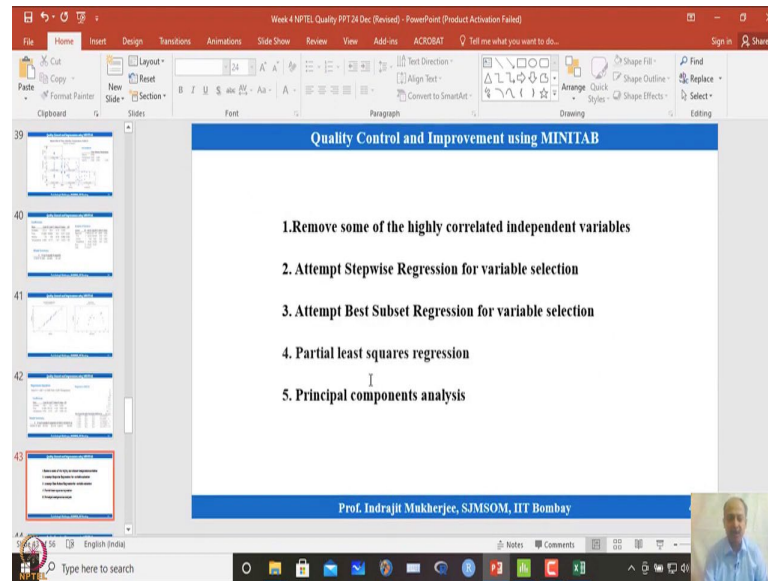
So, sometimes people can suggest why we should remove that one. So, there is a we can we can debate on that that which want to detain which one to remove like that. So, this is an art and this is not perfectly black and white scenarios like that in regression, at least multiple regression like that. But there are suggestions which can be incorporated like that. So, based on which we can select the variables..

So, one I have shown is that best subset regression is one of the option when we have different combinations of the variables and we can select one or two of them and then try to figure out which model is basically good.

Or we use stepwise regression and forget about everything of combinations like that, so whichever is the best will be the veiled and that model we will recommend like that, ok. But you should be careful about the model adequacy checks and all these things, ok.

Even if you have done stepwise regression also finally, you have to make a check of model adequacy over there, ok. So, that is the suggestions and there are other ways of dealing with multicollinearity coordinating which is more statistically sound like that.

(Refer Slide Time: 32:37)



So, one is partially square regression and one is principle component analysis based regression like that, ok. So, these things can be adopted.

So, we will stop here and we will we will try to discuss about another example where the multiple regression fails like that and error assumption fails and in that case how we have to deal with that. That is not discussed. It is discussed in simple regression. So, we will start from here. And another example I have on this time velocity temperature and yield and selection of the variables over here also we will discuss..

And then, we will move into the core concept which is the improvement phase and that is design of experiment. We will try to emphasize now on design of experiments and how do we how do you do design of experiments what are the things. So, basic idea of design of experiments in our next session basically.

Thank you for listening.