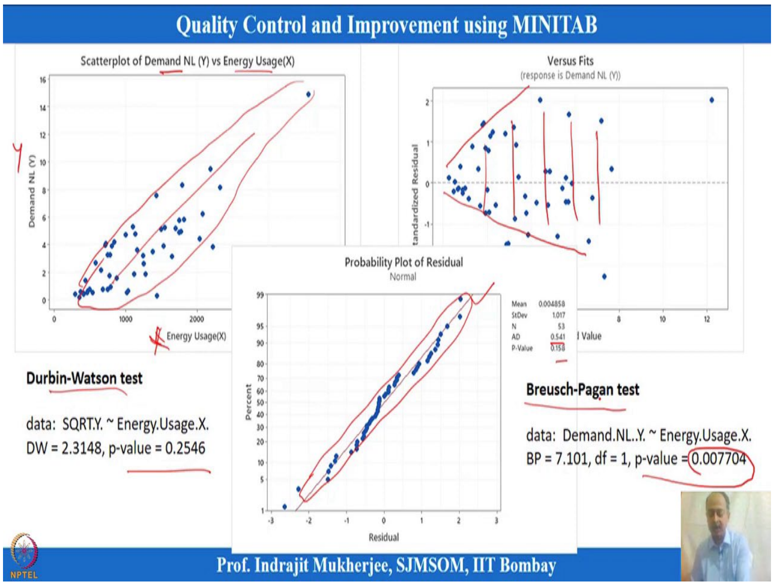
Quality Control and Improvement with MINITAB
Prof. Indrajit Mukherjee
Shailesh J. Mehta School of Management
Indian Institute of Technology, Bombay

Lecture - 25
Linear Regression (Continued) and Multiple Regression

Hello and welcome to session 25 in the course Quality Control and Improvement with MINITAB. I am Professor Indrajit Mukherjee from Shailesh J. Mehta School of Management IIT Bombay. So, previously we what we are doing is that last session, what we have done is that we are trying to understand regression simple regression and a model adequacy test like that.

So, there can be scenarios where model adequacy can fail and in that case what is to be done there also we want to see in regression. So, I will take one more examples to understanding simple regression, what are the complexities that can arise ok.

(Refer Slide Time: 00:51)



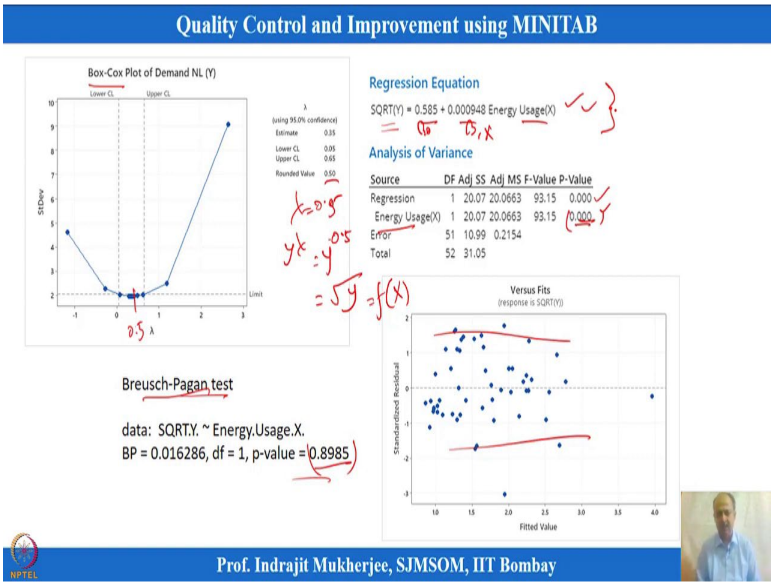
So, this is one of the examples that we are taking over here, where we have two variables Demand and one is Energy Usage, one is Y and one is X over here and this is the scatterplot that you see over here. And this scatterplot says that there is a linear relationship that exists between X and Y over here; and this is the X variable and this is a Y variable over here.

And what we are seeing is that, normal probability plot also says that there may not be any problem and Anderson-Darling test also says there is no problem, but when the residual what we are doing is the residual with fit for what we are seeing is that, there is certainly increase in variability of the residuals as we move ahead with the fits fit values over here. So, it indicates that although we have fitted a regression model over here scatterplot is finely prominent over here.

And we have done model adequacy checks and one of the model adequacy checks is normality distribution normal distributions of the residual, and it is not violating that the Anderson-Darling test shows that Durbin-Watson test also shows that the p-value is quite not significant.

So, in that case there is no problem in autocorrelation of the residual, but Breusch-Pagan test when I am doing this in R what I am saying is that p-value is quite significant that means, 0.0007 and that means, heteroscedasticity is an issue over here. So that means, the model cannot be generalized and we need to correct we need to correct this model.

(Refer Slide Time: 02:15)



And how do we do that? Then what can be done is that we already know that there is a transformation which can be used and we have used a Box-Cox transformation over here, which indicates a value of 0.05 approximately 0.5. So, lambda rounded value is 0.5 so, lambda equals to 0.5 and this indicates that there is a if there is y values, y to the power lambda we have to do and this will be 0.5 on this power over here.

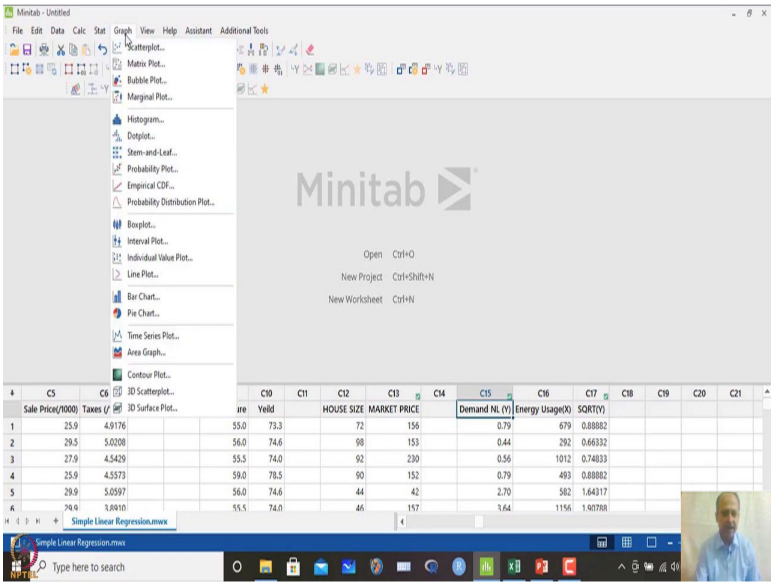
So, this is nothing but square root of y and this will be a function of x we which we have to generate like that ok. So, we have done that correction and after doing the correction what happens regression equation is square root of y and this is the β_0, that was generated and this is the β_1 or slope multiplied by x.

Analysis of variance also shows that the x variable is quite significant and the p-value is less than 0.05 and also when we plotted the residual versus fit over here, we do not see any abnormalities now over here.

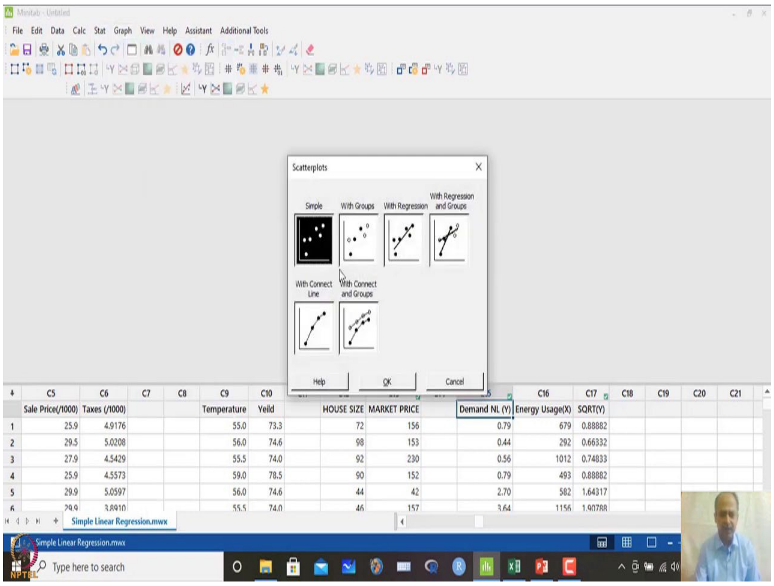
And the Breusch-Pagan test was again reconducted with this data set of residuals that was saved after we have generated the equation with square root of y and the value of p is 0.89, which is showing that the heteroscedasticity problem is not there.

So, this equation that we are generated can be used for a any unknown observation of x to predict what will be the y like that. We will generate square root of y so, that can be converted into y basically ok. So, this is one of the example. So, let us try to see how we have done this in minute I have so just to for you to facilitate.

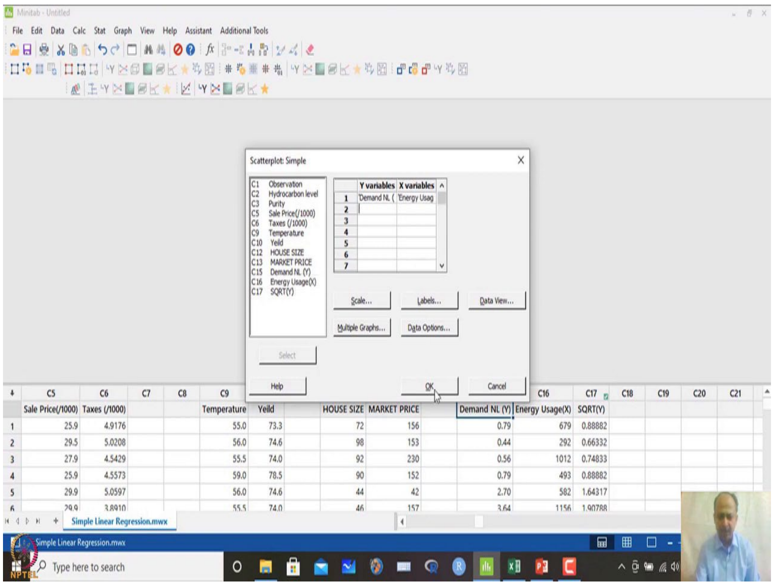
(Refer Slide Time: 03:39)



(Refer Slide Time: 03:47)

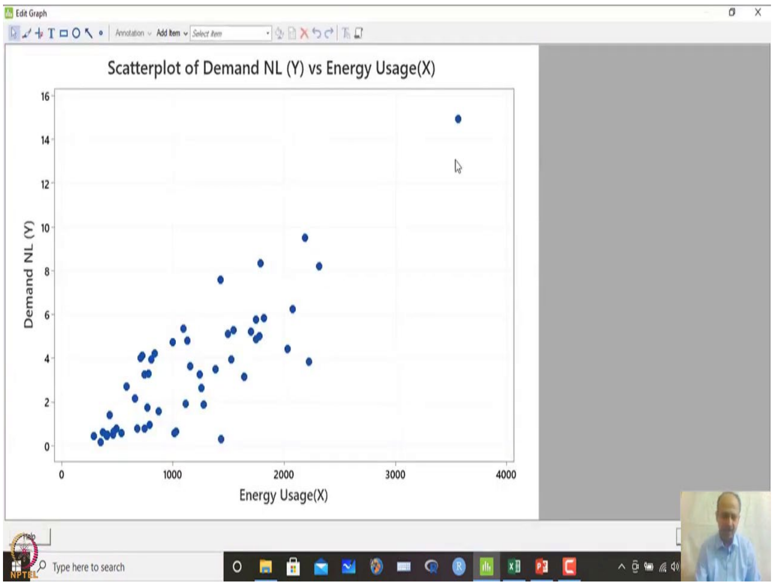


(Refer Slide Time: 03:50)



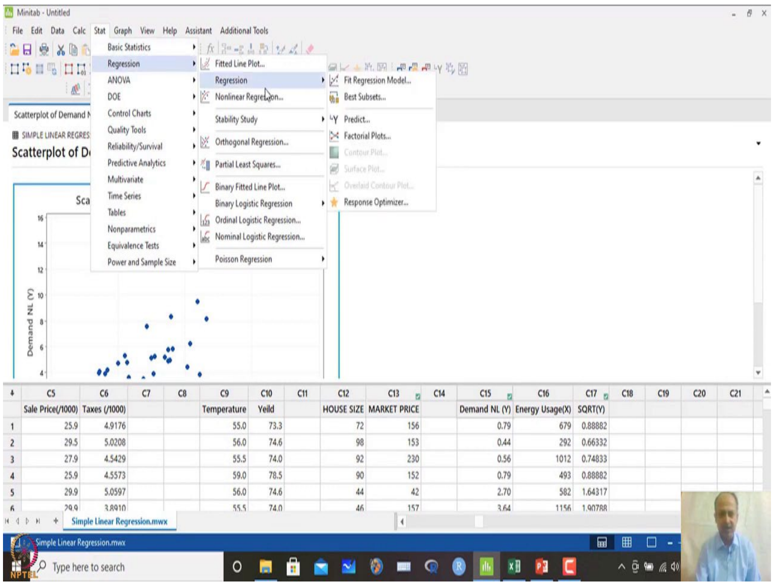
And so, here is the C15 and C16 column is given over here and what we will do is that, we will apply first we will see the scatter plot. So, graphically we can do the scatter plot over here and we can take let us say this is demand information and it is already taken over energy research over here.

(Refer Slide Time: 03:58)

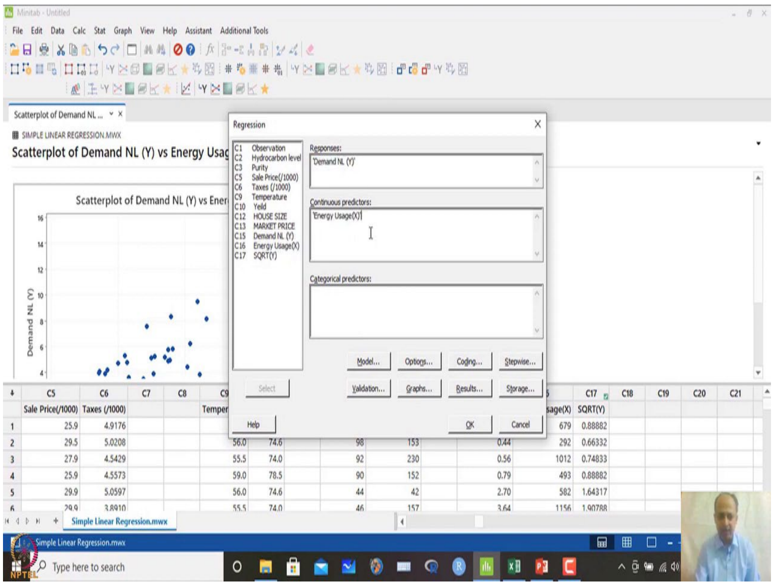


So, if you click ok Y and X variables, the same graph that I have shown in the excel sheets in the PPT this is the same graph what we see. So, strong positive relationship is exist over here that is shown in the scatterplot.

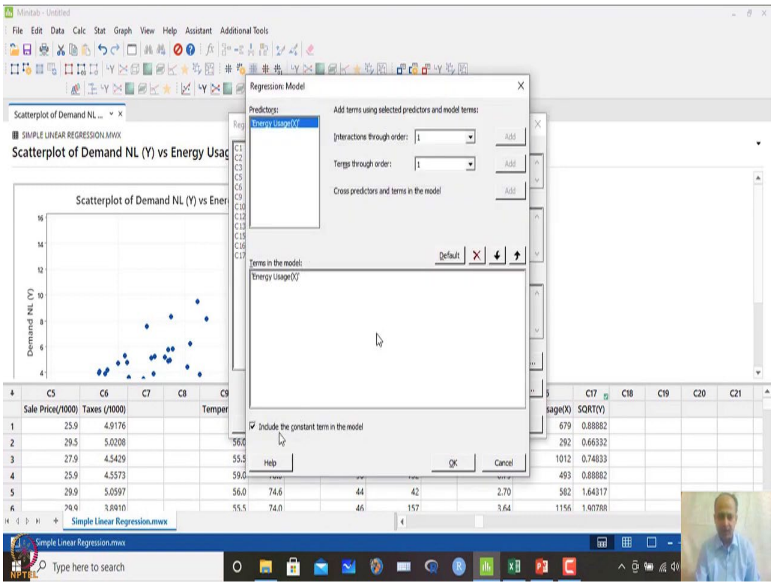
(Refer Slide Time: 04:11)



(Refer Slide Time: 04:14)

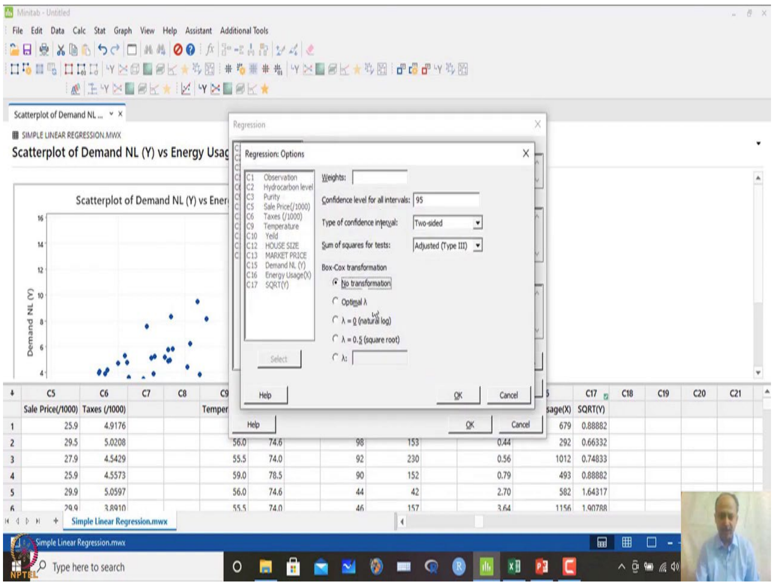


(Refer Slide Time: 04:23)

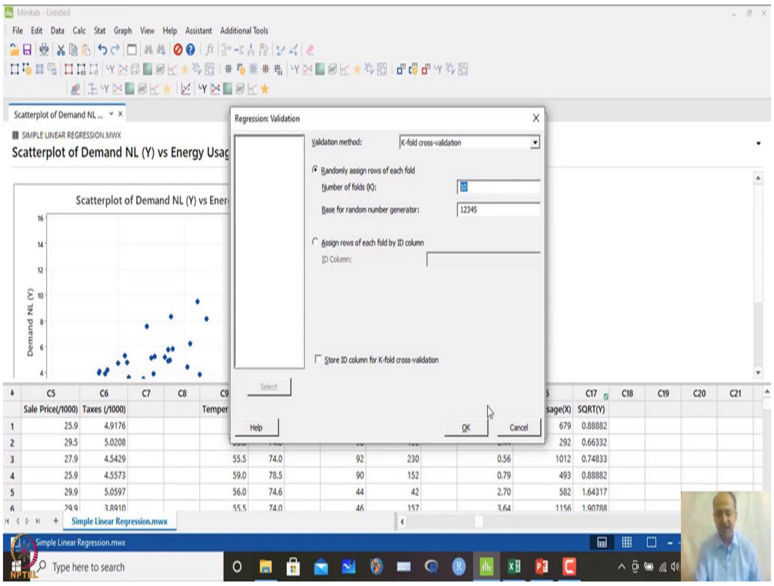


Then what we will do is that I will go to regression and then fit regression over here. So, fit regression model and then, we will do the demand and then we will take the energy usage over here and then in models what we do is that we include the constant term this is important. So, β_0 will be included in the term.

(Refer Slide Time: 04:30)



(Refer Slide Time: 04:45)

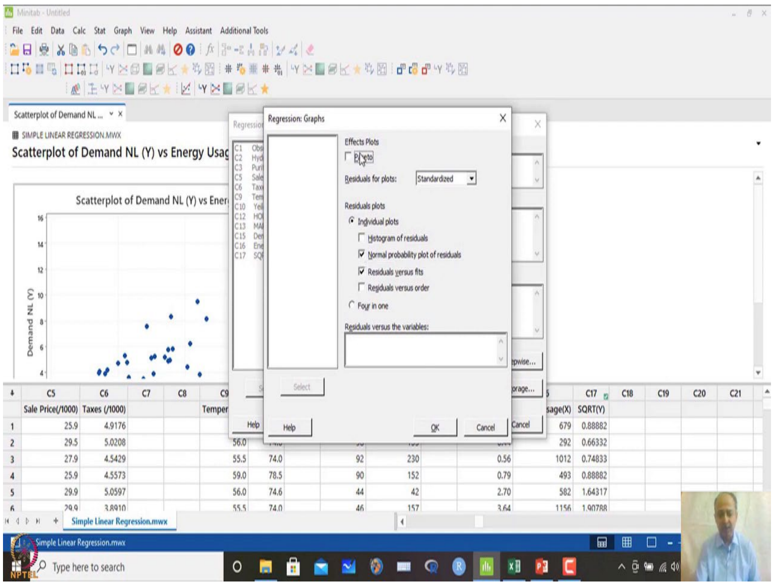


And in options at present no transformation is required. So, I have click to no transformation over here and other things we do not; we do not do anything over here, one more thing that can be done when whenever we are doing a regression over here. So, maybe validation so, this part can be done in many situation we do that in regression validating the models like that.

So, there are two methods over here, validating with proportional test sets and k fold cross-validations like that. So, people prefers to do K fold cross-validations so, and generally number of folds that is taken is 10. So, you can see cross-validation, how people are doing 10 fold cross-validation.

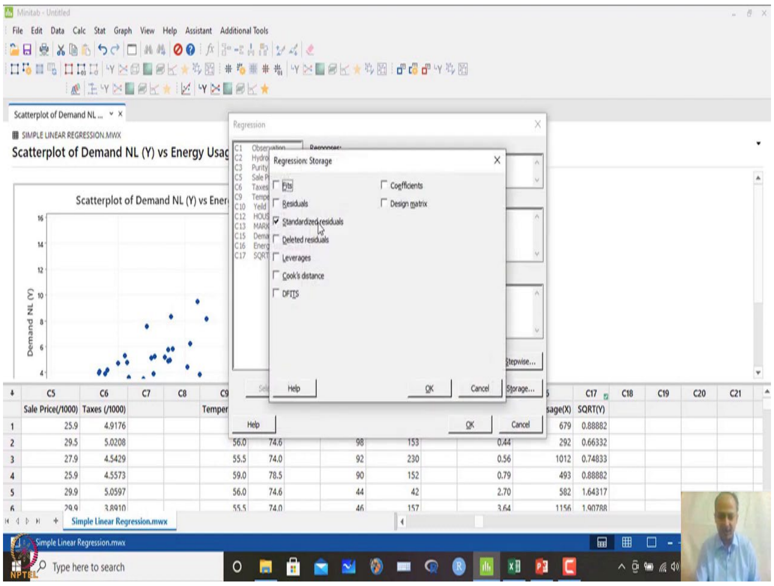
The theory behind this is simple very simple, I divide into 10 datasets like that one of the data set will be used as a test data set on which the r square value will be generated and that is the way we do cross-validation. So, that is one option we can keep in mind, when we are generating. So, that we can generalize so, but model adequacy test is required and that will show whether everything is fine.

(Refer Slide Time: 05:29)

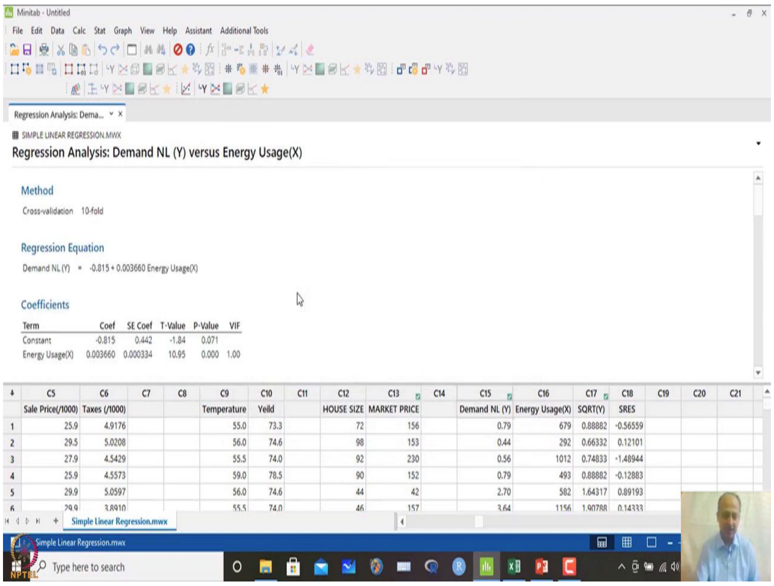


So, normal probability plot in and also what we can do is that there is a Pareto plot which can be done and residual what we want is standardized residual; because, we are talking about using standardized residual for normal plots and also for residual versus fit plot, and residual versus order plots can also be seen which says whether there is any dependency between the errors like that.

(Refer Slide Time: 05:50)

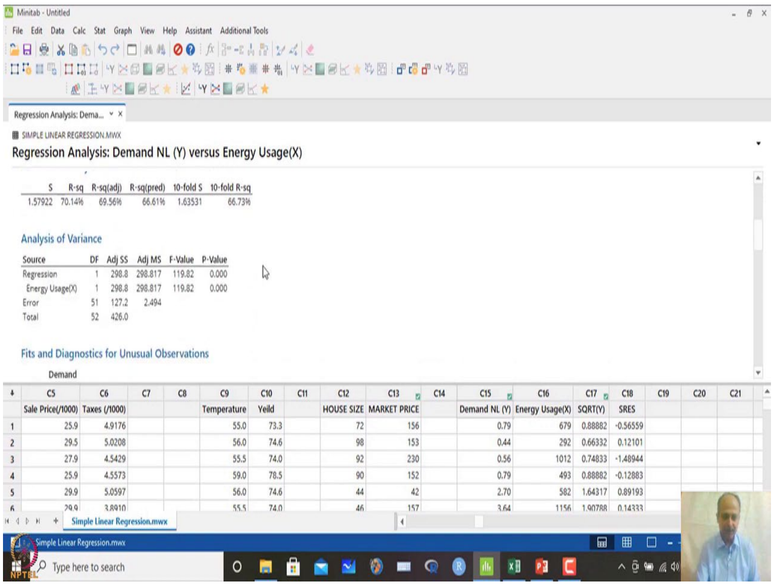


(Refer Slide Time: 05:56)

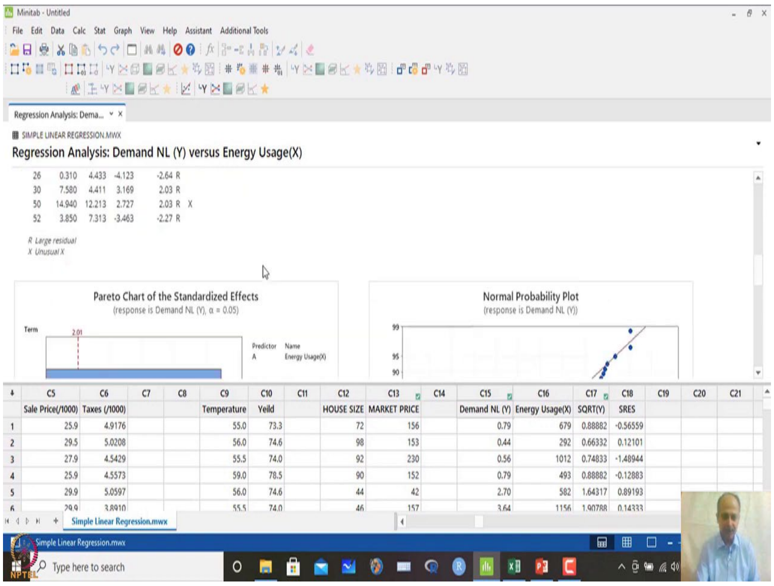


So, this can be seen and storage at finally, we can store the standardized residual and click ok and when you click ok over here what will happen is that you will generate equations also.

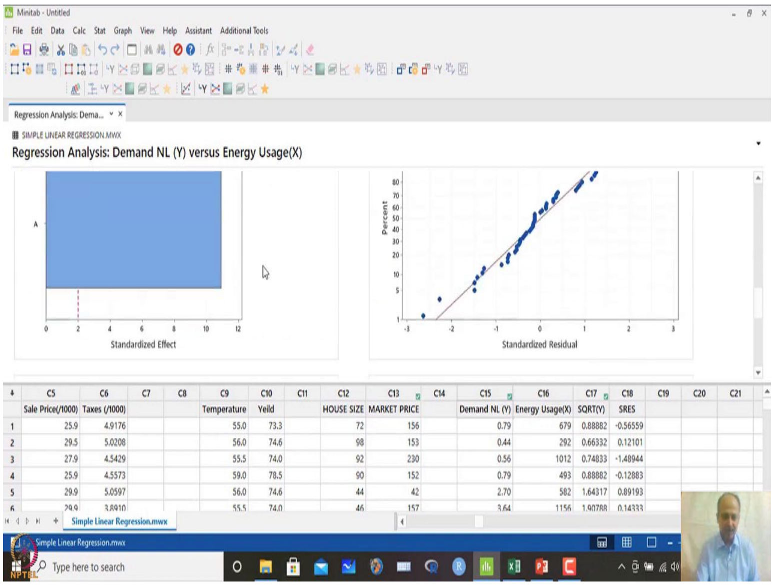
(Refer Slide Time: 05:58)



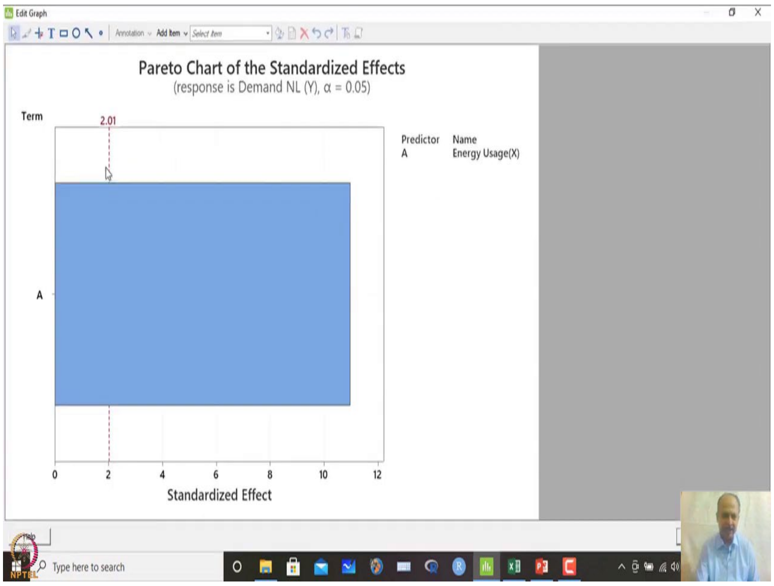
(Refer Slide Time: 05:59)



(Refer Slide Time: 06:00)

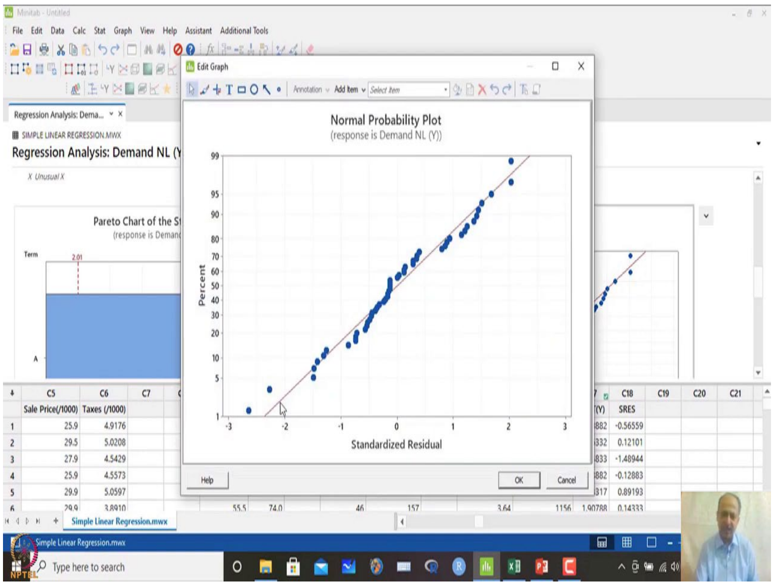


(Refer Slide Time: 06:03)

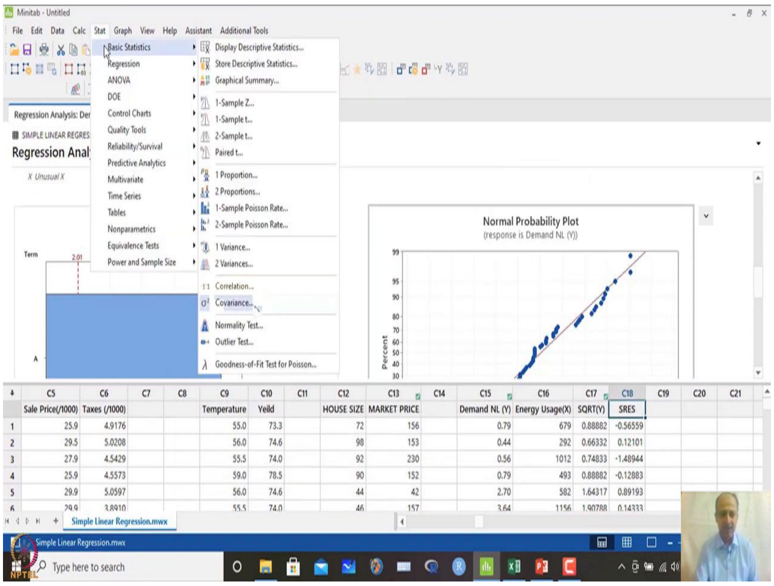


Then, you will generate these Pareto effects shows that anything beyond this red line indicates that that variable is significant that means, energy is said significant over here. So, it depends on the alpha value that we have taken. So, formula is there to find out this cut-off over here anything beyond the cut-off indicates that that variable is important. So, this is a standardized effect plot over here.

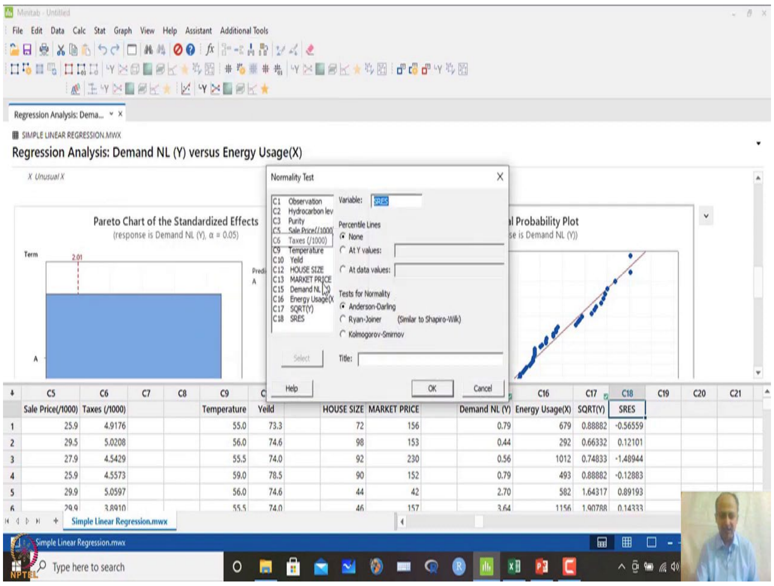
(Refer Slide Time: 06:20)



(Refer Slide Time: 06:27)

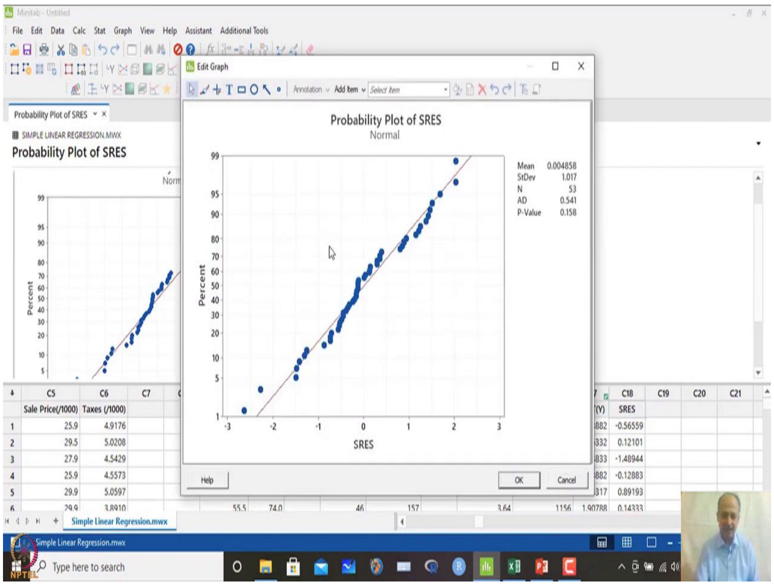


(Refer Slide Time: 06:30)



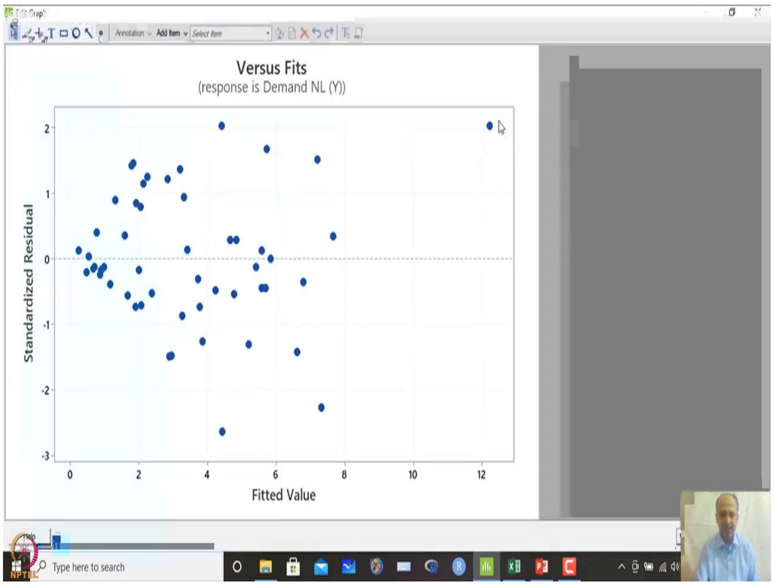
Normal probability pot that you see indicates that not much deviation is there, we have also save the residual. So, residual is saved over here so, we can check whether the basic assumptions of normality for the residual is ok.

(Refer Slide Time: 06:36)



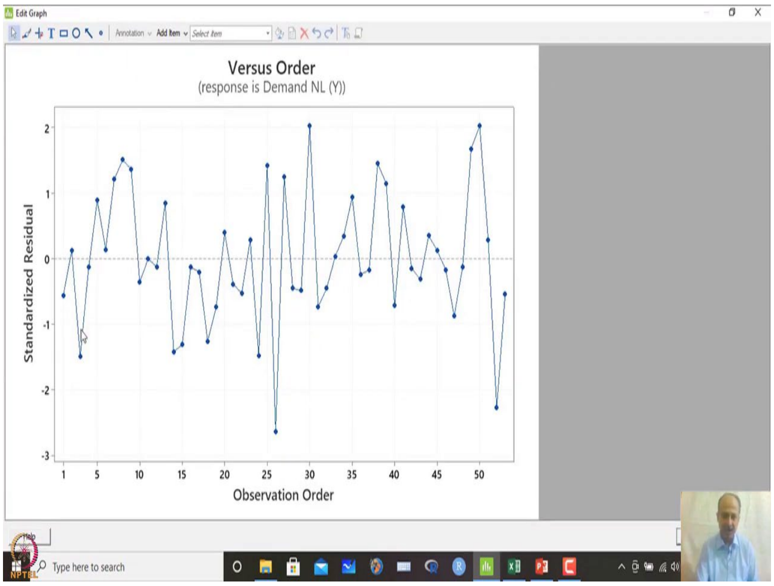
So, we can take the residual and check the Anderson-Darling test we can do that and what we see is that Anderson-Darling test is not showing any adversities over here.

(Refer Slide Time: 06:45)



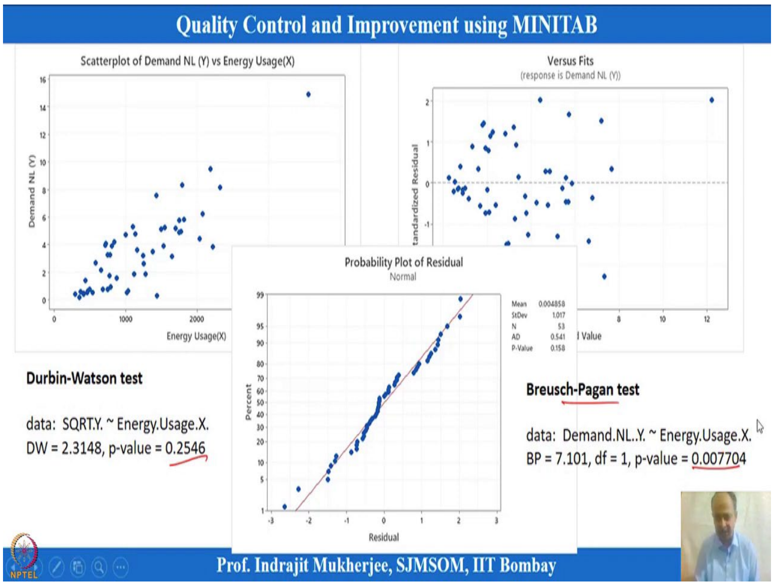
And then, here in the previous diagram what we have not seen is this one is that funnel shape what we told heteroscedasticity is prominent that we can see from this graph of residual versus fit.

(Refer Slide Time: 06:56)



And however, the autocorrelation aspects we do not see like there any trends we would would not see over here. So, in this case also we have done Durbin-Watson stat that I told that in this was done earlier, and that was not the problem that was not the issue that we have identified over.

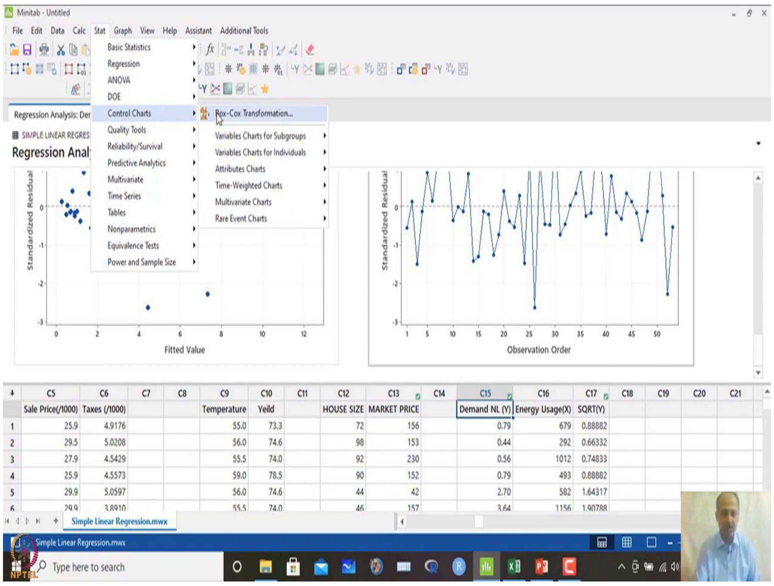
(Refer Slide Time: 07:08)



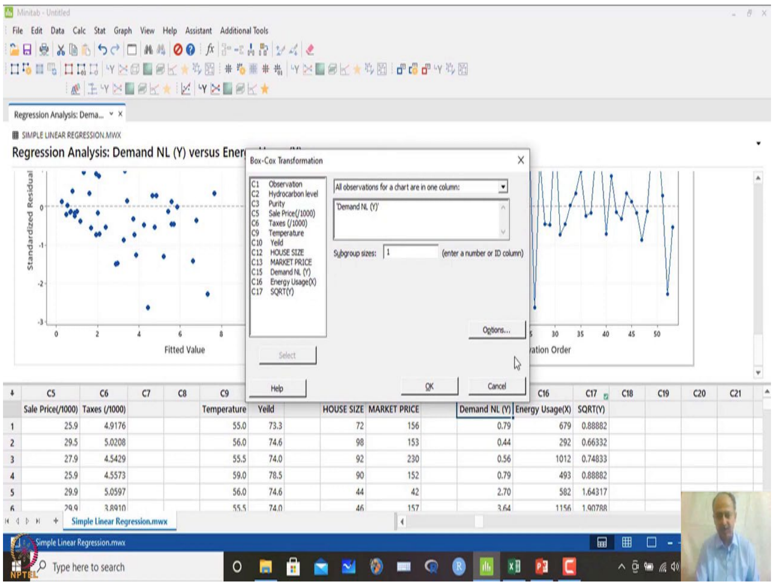
So, the Durbin-Watson autocorrelation is not an issue over here because, p-value is coming out to be, but Breusch-Pagan test showed that there is significant heteroscedasticity that is that we are getting. So, in that case model cannot be generalized.

So, we need to do something on this. So, what we can do is that? We can go for transformation, we can go for transformation over here and using the Box-Cox transformation this was done, and what we can do is that this is the then we have to convert the Y variable over here. So, that the residuals we will not have a model adequacy problem over here.

(Refer Slide Time: 07:51)

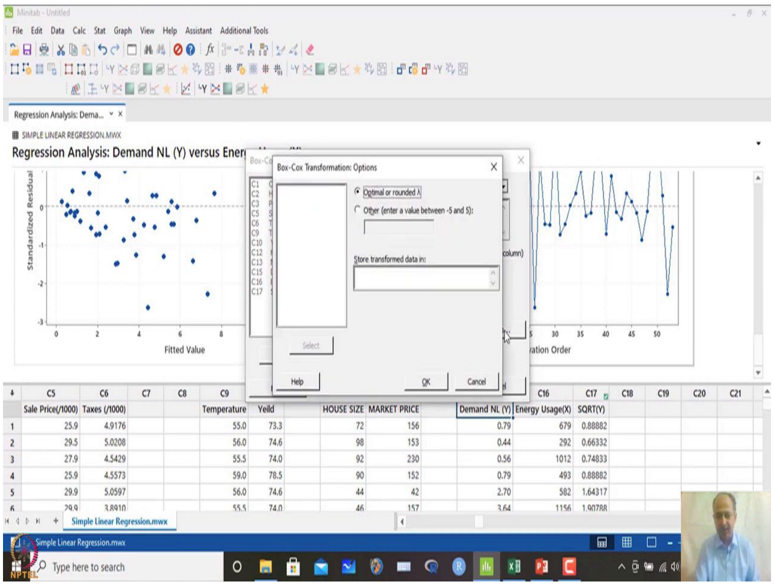


(Refer Slide Time: 07:53)

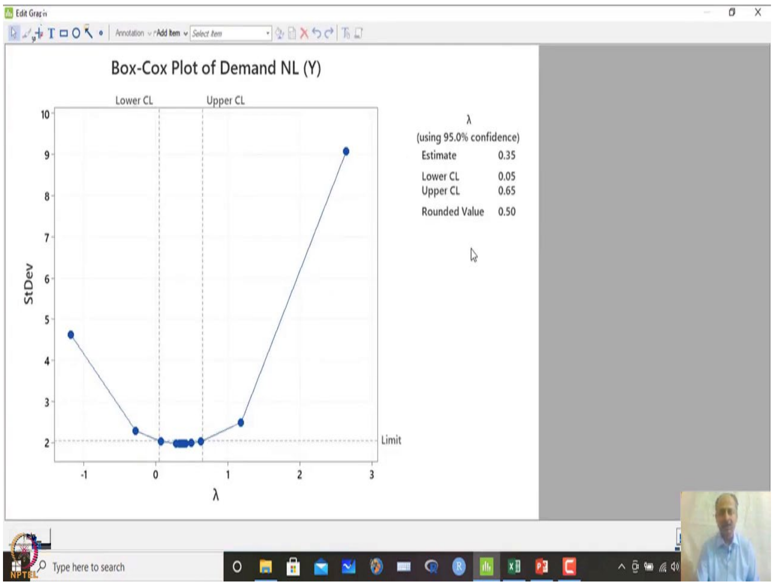


So, what we have done is that? We have gone for the first option may be that Box-Cox transformation. So, how do you do Box-Cox transformation? Box-Cox transformation of which variable demand over here subgroup size is this.

(Refer Slide Time: 07:59)



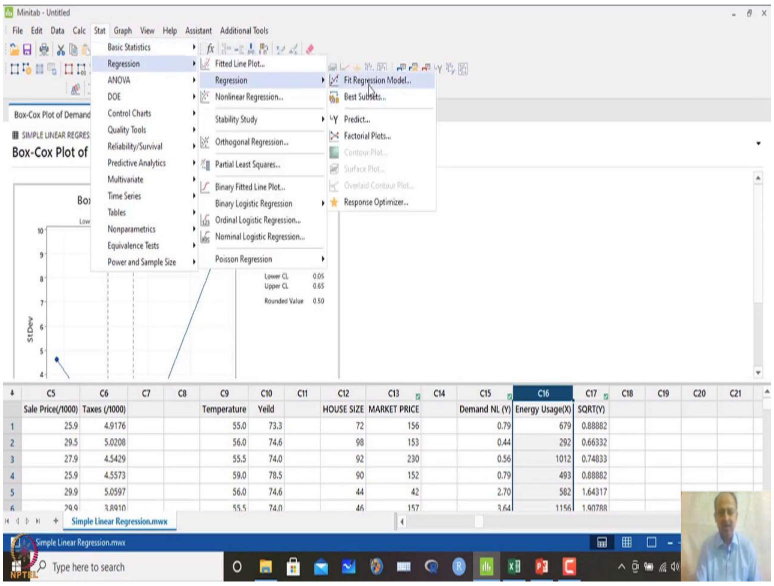
(Refer Slide Time: 08:04)



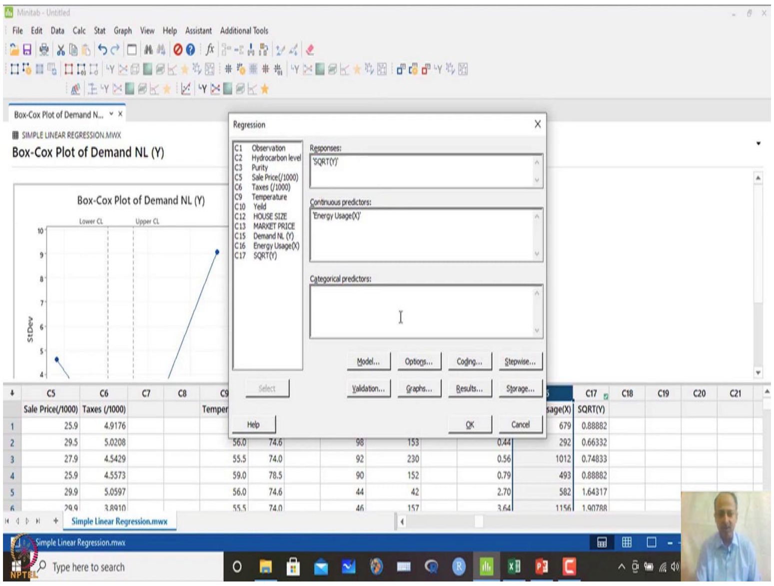
Where we want to save that one optimal or rounded value so, let us try to see first then we will save. So, in this case if we click ok what will happen is that Box-Cox will recommend you what is a transformation that is required. So, over here you see the estimated values 0.35, but you can round it off because lower and upper confidence level will include this.

So, 0.5 we can consider as rounded value, because lambda, Y to the power 0.5 is understood by many people that is square root transformation. So, the square root of Y is required so, we can use that. So, what I have done is that I have taken square root of Y over here, that is in C17 and then I have regressed C17 with C16 like that.

(Refer Slide Time: 08:38)

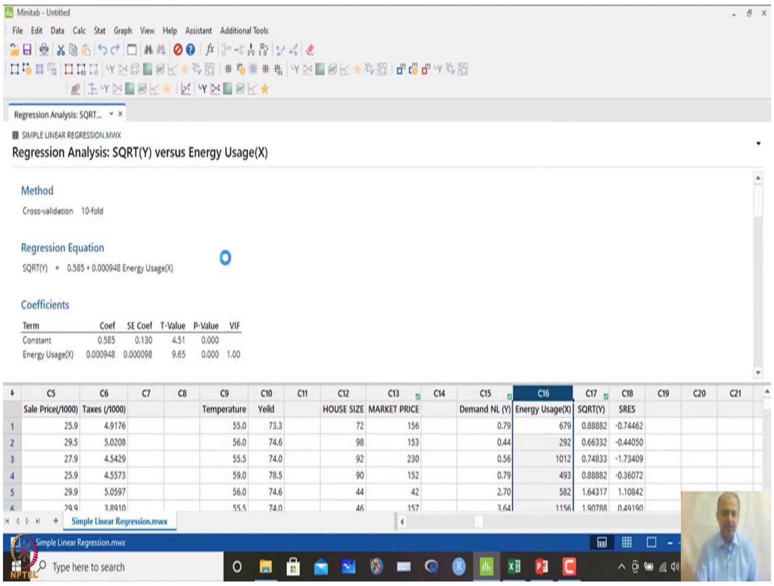


(Refer Slide Time: 08:41)

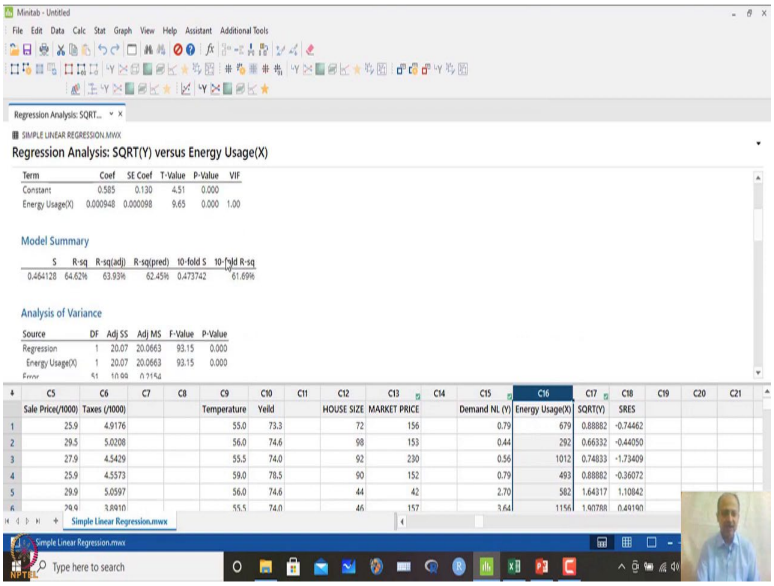


So, then what I have done is that regressed regression models over here fit regression model. So, everything remains same only instead of this I have taken square root of Y over here.

(Refer Slide Time: 08:47)

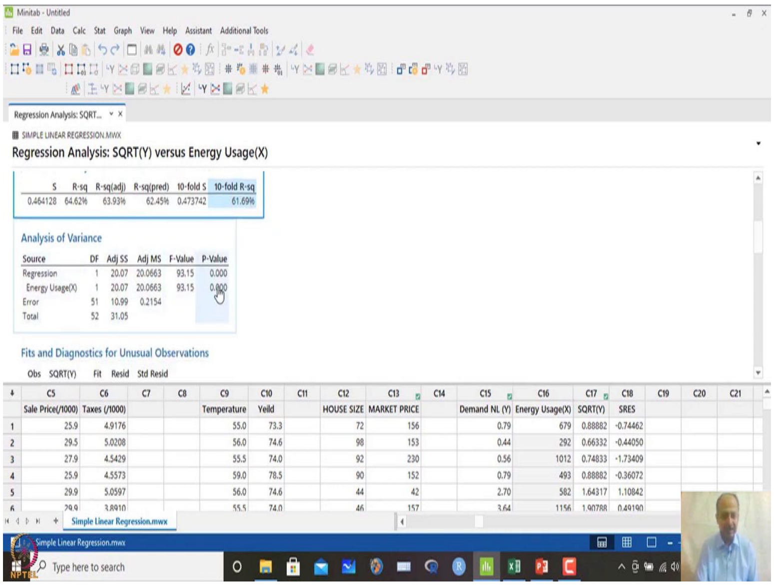


(Refer Slide Time: 08:48)

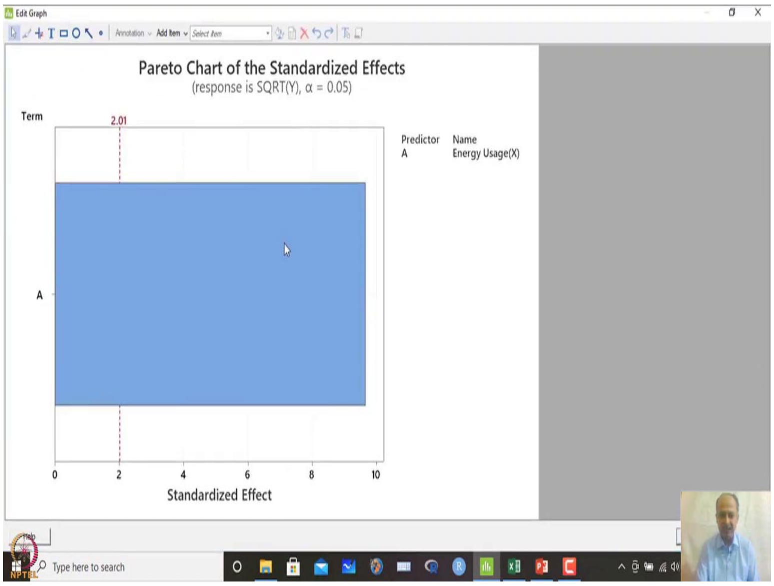


Everything remains same and I click ok and then what is generated is that, now energy usage is going to be significant p-values over here, even after transformation and 10 fold cross-validations these values is around 61 that means, cross-validation values that we are getting around 61 percent ok.

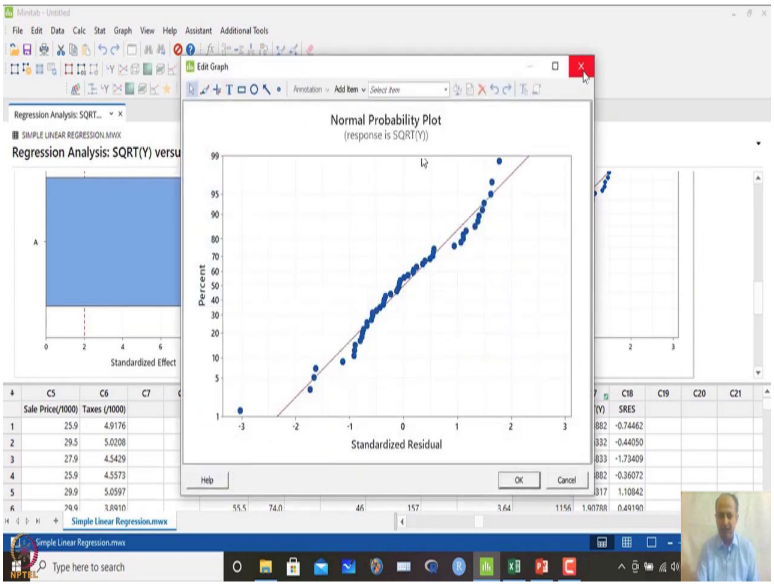
(Refer Slide Time: 09:08)



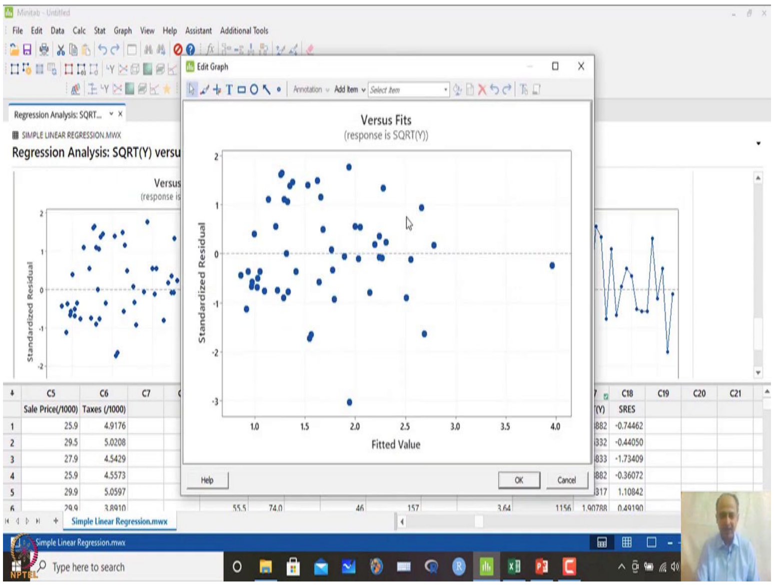
(Refer Slide Time: 09:15)



(Refer Slide Time: 09:23)



(Refer Slide Time: 09:31)



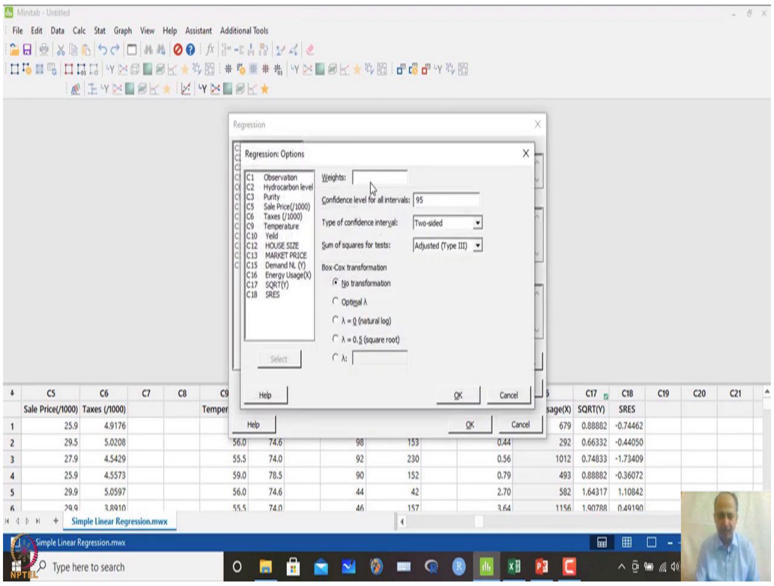
So, that is on the lower side, but still we may be satisfied with this energy but, there is a significant relationship that exists between this data set and here also this effect plot also shows that this is significant over here. Energy usage is prominent and this variable needs to be considered and there is no problem with the normal probability plot also.

And we want to check whether the final one is heteroscedasticity is problem is eliminated here also we do not see much problem in the behavior of the residuals with respect to X or with respect to fitted values like that. So, and we have done the testing, we have done the Durbin-Watson test for autocorrelation.

We have also done with the transform data, with the transform data and the residual that we have generated and we have also seen that the Breusch-Pagan test also does not show significance, that we that observation we have done when I have converted the data so over here.

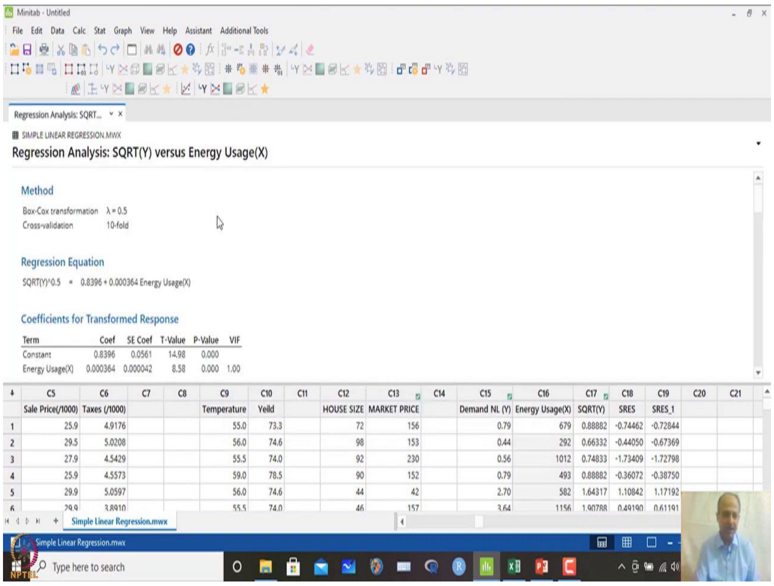
What you see is that? After square root transformation Breusch- Pagan test that this is 0.8985 and that is more than 0.05 and that indicates that the problem of heteroscedasticity is removed by using a box cox transformation over here, using a Box-Cox transformation over here. And MINITAB gives you an option that means so whenever I am doing regression over here. So, if I close this one and you can do the transformation.

(Refer Slide Time: 10:26)



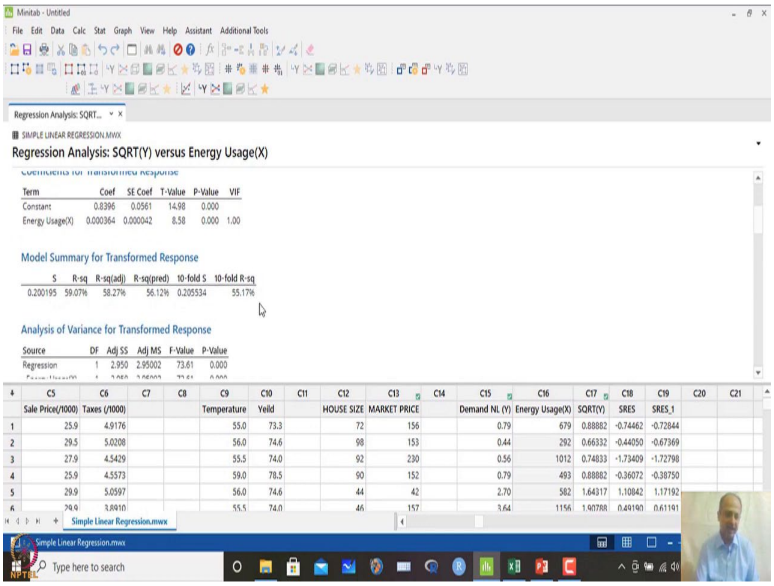
So, here regression what you can do is that regression fit regression model over here, and in options what you can do is that you can say transformation lambda equals to 0.5 transformation I want.

(Refer Slide Time: 10:32)

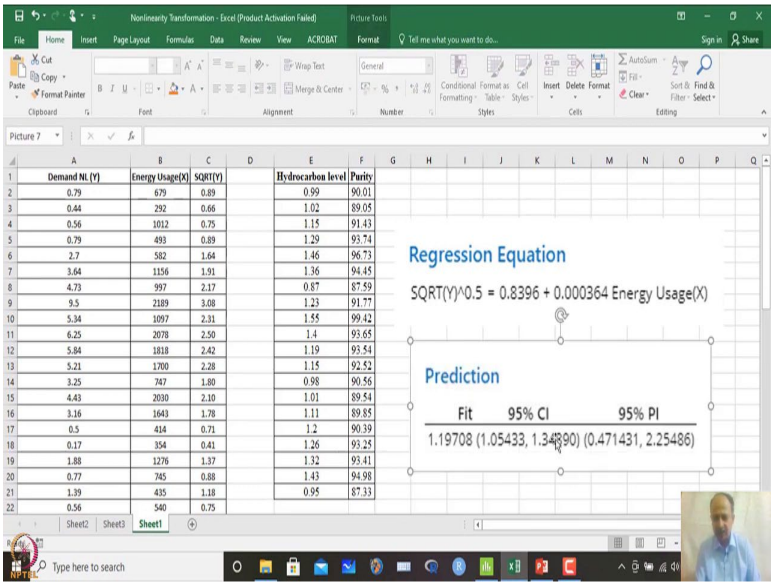


And after doing that you click ok. So, MINITAB will automatically generate a Box-Cox transformation with lambda equals to 0.5 cross-validation 10-fold cross-validation we are doing.

(Refer Slide Time: 10:40)



(Refer Slide Time: 10:48)

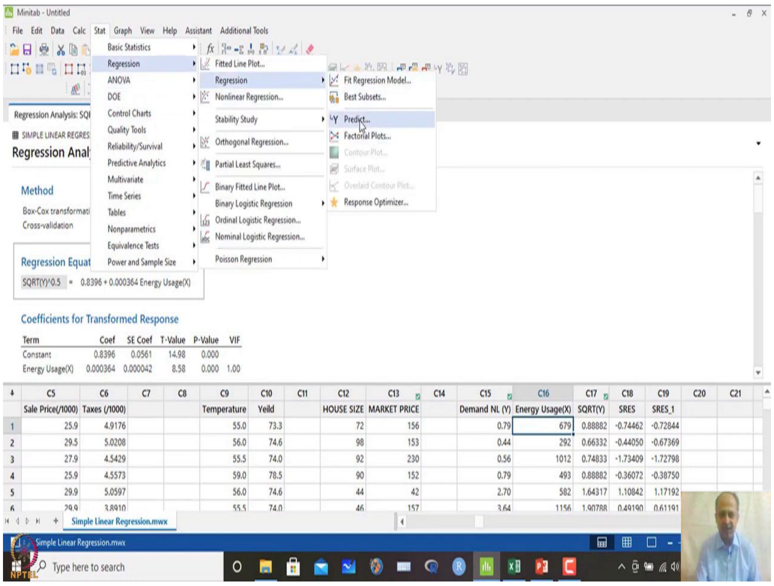


And in that case, what is a corresponding p-values, what is the final equation also it will show. So, this is the final equation it is showing. So, copy as picture we can just paste this one over here and you can see what is the equation final equation that is generated over here.

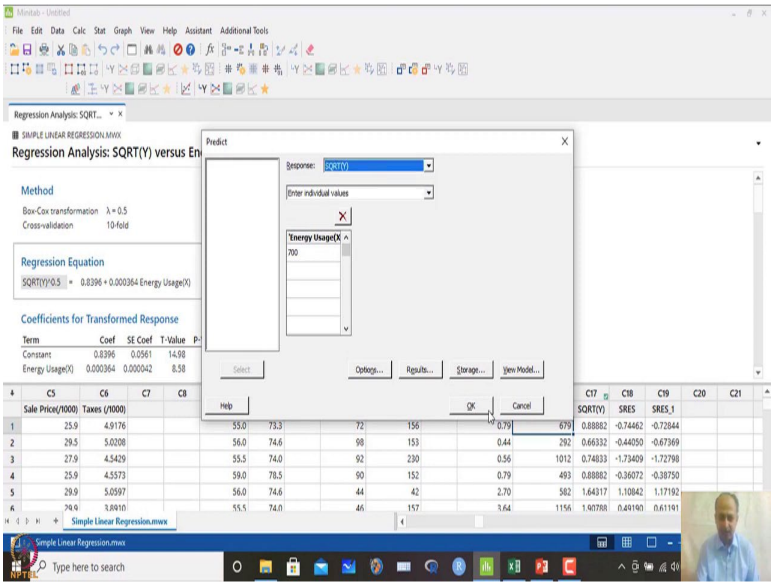
And the residual that will be generated that is a actual versus minus predicted values for a given value of X and all the residuals when we plot that one and we do the model adequacy check of heteroscedasticity and other things what we observe is that it will it is satisfying basic all conditions, all conditions it is satisfying. So, this regression equation can be used for generalization.

So, within the range of X so, we will not extra pull it, but within the domain of X where the equation is generated within that, any value of X you give I can predict what will be Y over here. So, in this case how do you do that? So, in this case let us assume that this I want to predict something that energy level 676, what will be around 700. Let us say 700 what will be the value of demand. So, what will be the value of demand?

(Refer Slide Time: 11:42)

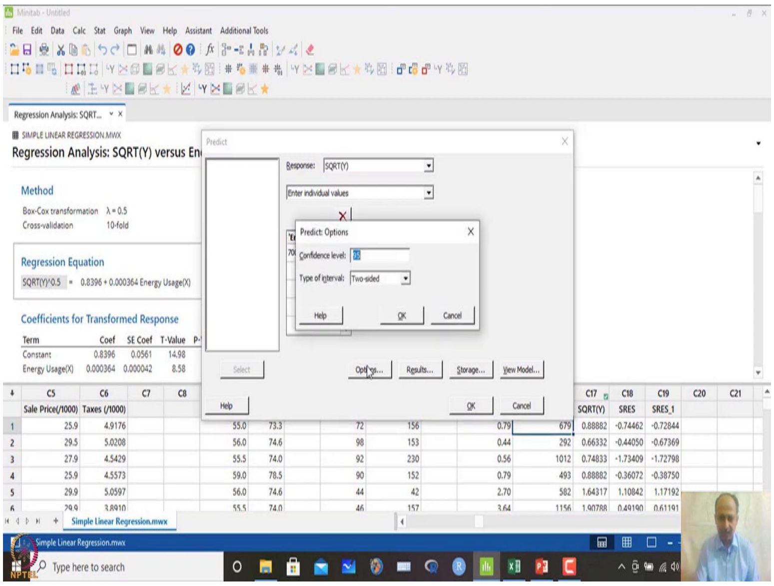


(Refer Slide Time: 11:46)



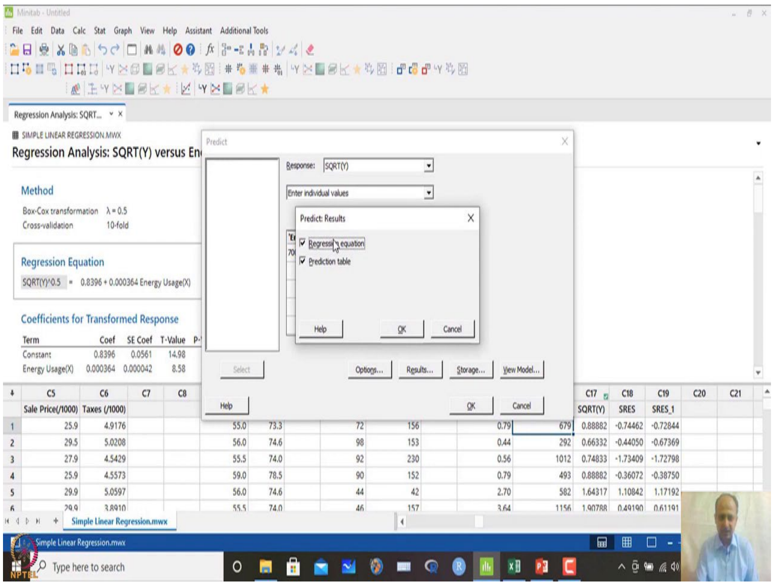
So, in this case what we will do is that regression and regression what you have is predict. So, you can do prediction over here. So, energy uses let us say 700 what will be the predictor square root of Y over here ok. So, we can also see this square root of Y over here.

(Refer Slide Time: 11:57)

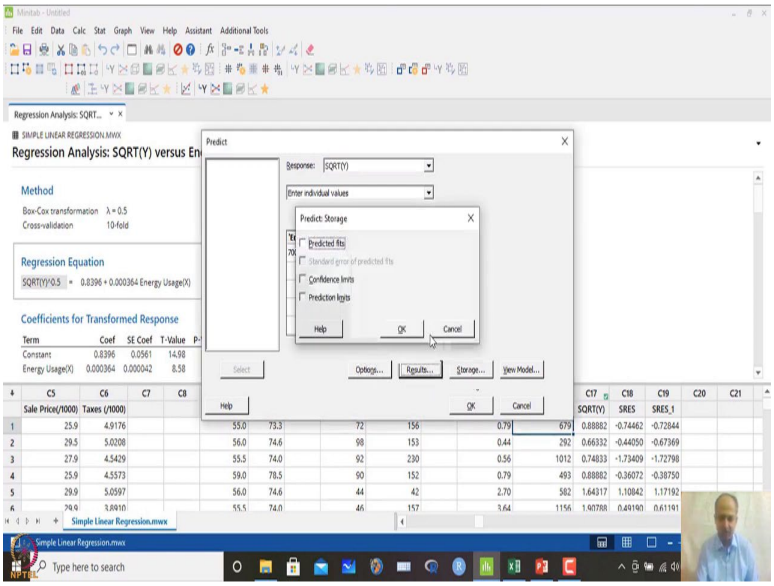


So, in this case options confidence interval that can be generated.

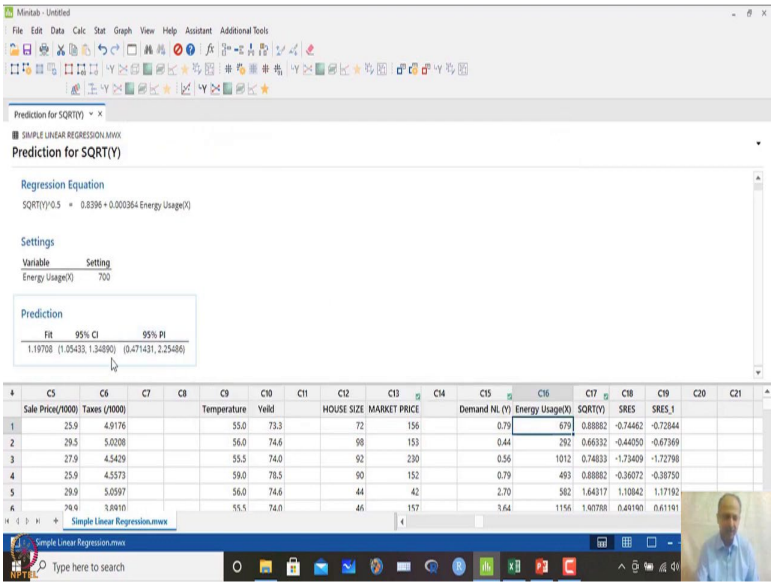
(Refer Slide Time: 12:01)



(Refer Slide Time: 12:05)



(Refer Slide Time: 12:09)



So, results over here, regression and predicted tables over here. So, in this case storage if you want to store that one that is also possible, new model is not required. So, in this case what will happen is that this predicted value will be given over here. So, if I copy this one and I paste it over here and I can paste it over here and to show you the values what is being generated over here from the regression equation.

So, fit value is around 1.97. So, this is approximately. So, this generate a square root of Y, this is square root if you make square of this you will get the actual value of Y over here. So, it will give you a confidence interval, it will give you also a prediction interval. So, this can be we can generate the confidence interval based on certain formulations over here.

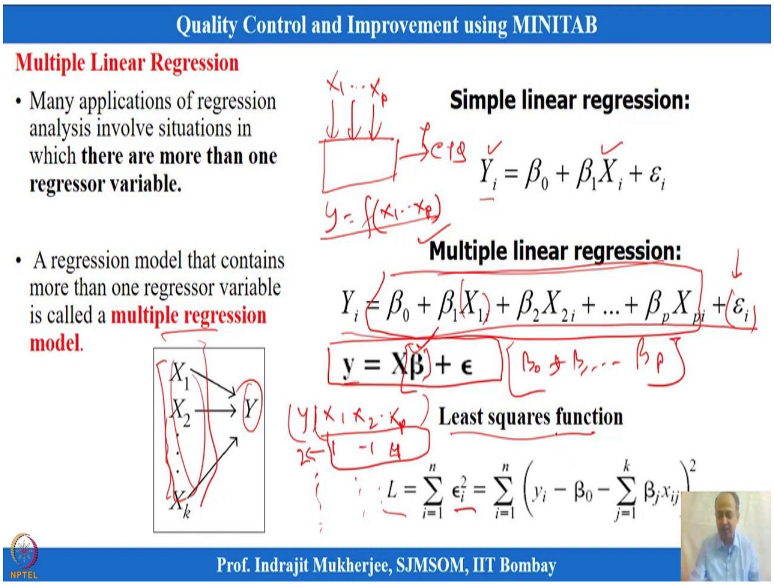
So, this is the expected value, but it will have a range it will not be exactly values that you that will come out, but the range of this confidence interval of the values that is predicted is approximately 1.05 to 1.34 and prediction interval is around 0.04 to 2.25 0.04 to 2.25. So, this is this can be also done in MINITAB software.

So, what is the prediction interval, what is the confidence interval for a given value of X? So, given value of X is given as 700 for this I want to predict like that.

So, 700 should be within the operating zone of that control variables that we want to predict, after generating the regression equation and we can do all the checks and finally, we will adopt the equations like that. In real life also we develop the models between Y as a function of X and then, try to use that to reach to the optimal solutions like that where should be the x so that I get the best Y like that ok.

So, this is a simple linear regression where we have many things to understand model adequacy cross-validation then R square values then, beta is significant or not slope is significant or not ANOVA analysis all these things needs to be considered when we are talking about regression ok. So, we can extend this concept of regression to multiple regression, simple regression to multiple regression over here.

(Refer Slide Time: 13:58)



So, multiple regression is nothing but when we have more than one X over here, this is the natural scenario over to we will encounter most of the time. So, this is the matrix of X and there will be one single Y over here. So, this is known as multiple regression, this is known as multiple linear regression.

So, simple is one Y and one X like that simple is one Y and one X, but in case of multiple linear regression what will happen is that there will be multiple variables that will influence that means, in a process when we are talking about this is the CTQ, which is y that can be influenced by many X over here.

So, X p up to X p variables. So, X 1 up to X p variables can influence the process CTQ and all may be potential X all maybe potential X over here, then we want to generate the functional relationship between y and function of all X 1 to X p and I want to generate that function.

So, how do we do that? We do the that by regression, we do that by regression when I have more than one X variable we call it as multiple linear equation and this is the expression that is generalized expression that you see. This is the model empirical model functions that is given over here, and there will be some error whenever I am generating a function and it will not be realistic.

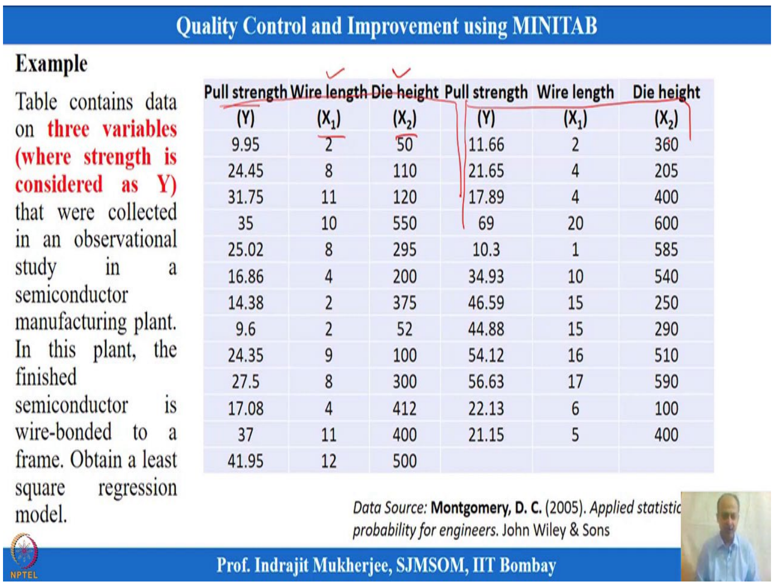
And so, there will be some difference between actual and predicted that will be error or residual and that will be saved over here. So, that we can see and this is the matrix notation that you see in case when multiple X and multiple we have multiple X for predicting a single Y like that. And this is the beta coefficients that is β_0 plus β_1 like this up to beta p. So, this is estimated over here. So, this is a matrix that you see over here and this beta can be estimated from these values of Y and X.

So, I will have y and I will have multiple observations X_1 up to X p observation. So, this we will have some values let us say for a given value of this can be some values 2 over here, and this can have 1 minus 1 4 or something like that. So, there will be different value. So, 1 row of X will give me certain y over here. So, this is the setting condition that will give me the process condition output condition over here that is the given over here.

So, if you have multiple y observation and multiple X observation and for any row of X we will have y conditions like that then, that will if that is there in that case by matrix we can matrix formulation, we what we can do is that matrix algebra what we can do is that, we can generate what is the value beta over here.

And this is based on some least squares functions over here that is error minimization. Basically, it will minimize error defined partial derivative should be taken over here and that will give me the values of β_0 up to β_p ok.

(Refer Slide Time: 16:36)



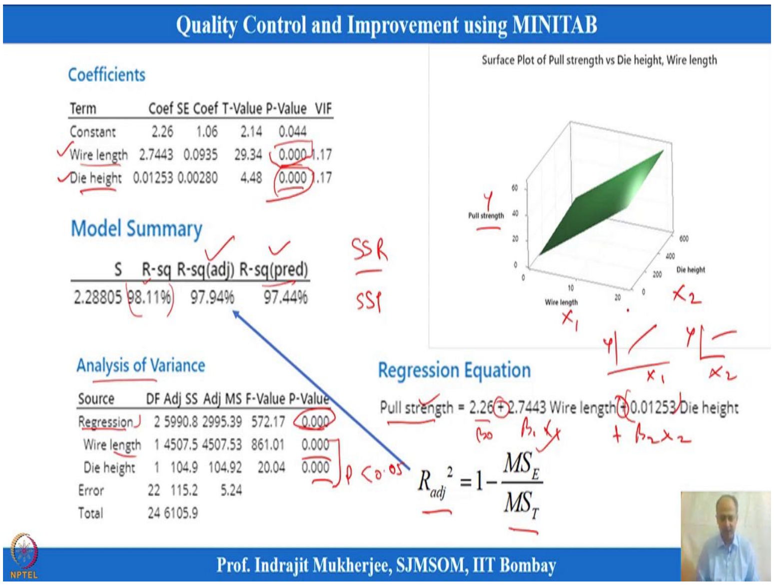
Pull strength (Y)	Wire length (X_1)	Die height (X_2)	Pull strength (Y)	Wire length (X_1)	Die height (X_2)
9.95	2	50	11.66	2	360
24.45	8	110	21.65	4	205
31.75	11	120	17.89	4	400
35	10	550	69	20	600
25.02	8	295	10.3	1	585
16.86	4	200	34.93	10	540
14.38	2	375	46.59	15	250
9.6	2	52	44.88	15	290
24.35	9	100	54.12	16	510
27.5	8	300	56.63	17	590
17.08	4	412	22.13	6	100
37	11	400	21.15	5	400
41.95	12	500			

And MINITAB does it automatically for you. So, in this case one of the example that I am taking over here is known as pull strength and I want to see whether there is any relationship between wire length and die height over here. These are the two X variables and I have multiple observations over here, and this is one set of observation that you are seeing and this is another set of observation that you are seeing over here.

Same variables and same Y over here so, we have placed it side by side. So, that you can see the complete data set like that there are two variables or predictors and one predicted values or CTQ's over here, and I want to see whether the both the variables are important to be included or one is sufficient to model is one and which way I should develop the model.

So, this is a scenario, where multiple regression is required to so, that we can model Y as a function of multiple X; X1, X2 over here.

(Refer Slide Time: 17:29)



So, what we do over here is basically we again do the regression analysis here we instead of one variable two variable information you will get wire length and die heights over here, and this is P-value is showing to be significant again the same interpretation. If P is significant in that case, it will indicate that these variables are important and this should be included in the models like that ok.

And R square what you can see the interpretation that we have seen R square that is how much of the variability SS regression by SS total how much of the variability is explained and this is coming out to be 98 which is quite significant and, but what we do is in multiple regression we see R square adjusted ok.

So, R square adjusted gives you a precise estimation over here which is formulation is given over here and we go by mean square, we go by mean square calculation because this prevents overfitting of the model. This is preventing overfitting of the model because, if you keep on adding X R square will increase, but it will not increase R square adjusted until and unless it has significant influence to reduce the mean square error like that ok.

So, another is predicted value that means one observation will be dropped and what is the value prediction, how close is that prediction over here. So, that can be done by R square predicted value over here and MINITAB gives you all options to see how they are calculating R square predicted and you can see some resource on out books resource how R square predicted is.

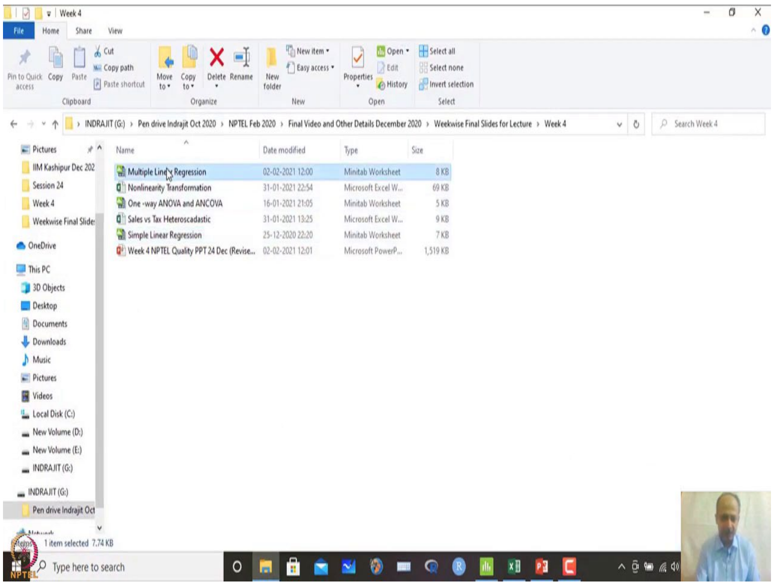
So, each of the observation will be removed, it will generate a equation and then predict the single observation that is the way we do for R square prediction over here.

Analysis of variance then analysis of variance shows that when I have two variables over here overall it is significant on that regression equation yes it is significant out of these two variable, which is important both are important because the p-value is less than 0.05. So, our interpretation is that we should include both of them in the model ok.

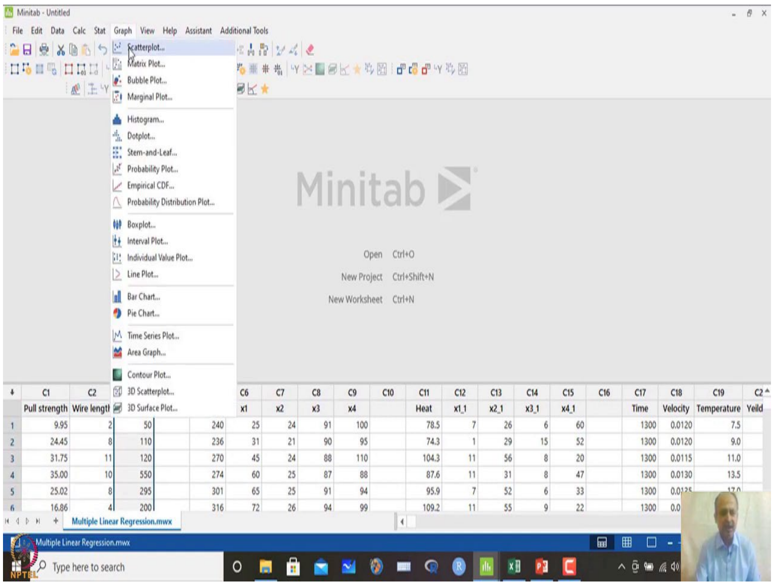
And the overall equation that you will see is that expected value of strength is equal to this is β_0 over here, and equation is generated by MINITAB $\beta_1 X_1$ this is the equation and this is plus $\beta_2 x_2$ over here, both are significant this coefficient are significant and this is plus sign this plus sign indicates that both are linearly positively related over here. X1 increases, y increases, X2 increases, y also increases like that.

So, these things can be interpreted out of the analysis and how we do that we want to see and this is the surface plot what you see over here, and this is showing you that pull strength that is y how it is related with X1 and the MINITAB also gives you an option to make a 3D plots like that. If you have only two X variables and one Y variables then plotting is also possible like that.

(Refer Slide Time: 20:12)

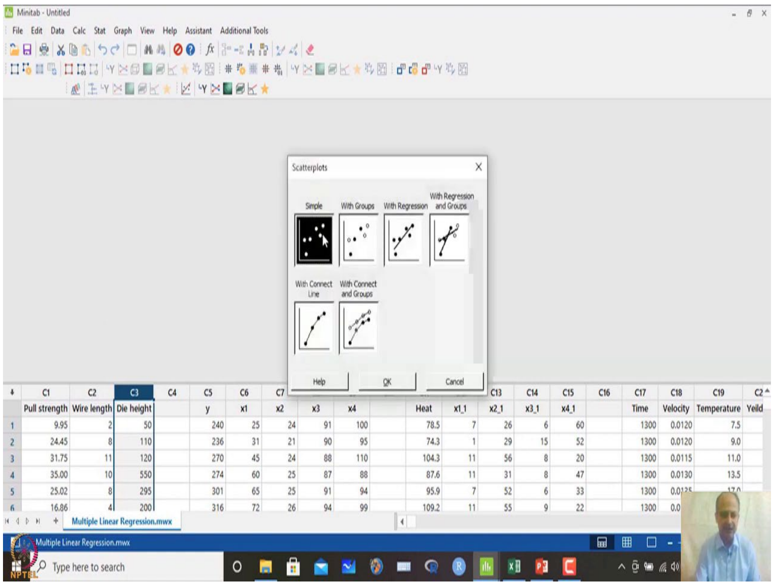


(Refer Slide Time: 20:21)

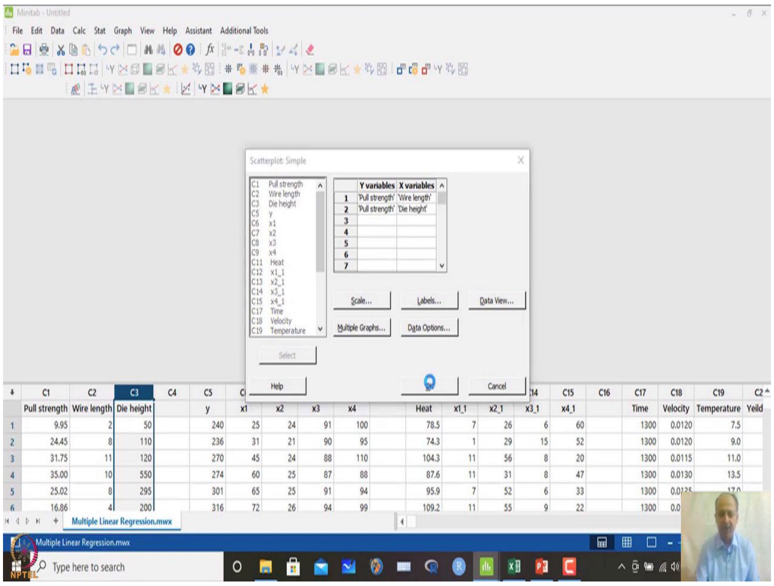


So, we will show that one also. So, in this case how this data is generated later let us go to the analysis part of this. And so, what we will do is that, we will go to the data set and we will try to see multiple regression and when we open the data set this is the data set that is C1 to C3 columns what you observe over here pull strength, wire length, and die height.

(Refer Slide Time: 20:29)

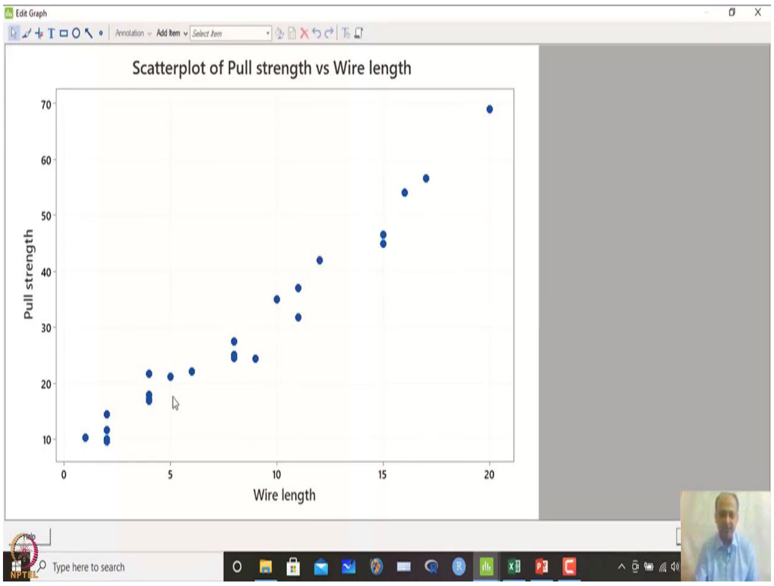


(Refer Slide Time: 20:36)

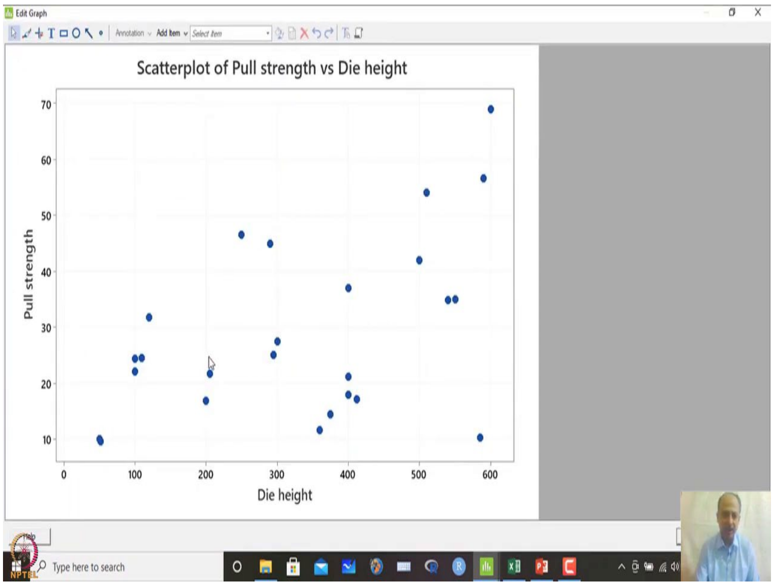


So, what we will do is that, stat we can see graphically first how these two variables are and each of them are related to the Y variable. So, what we will do is that we will say pull strength and wire length I want to see and again pull strength with die height I want to see, how the relationship is and it will plot two diagrams like that.

(Refer Slide Time: 20:43)



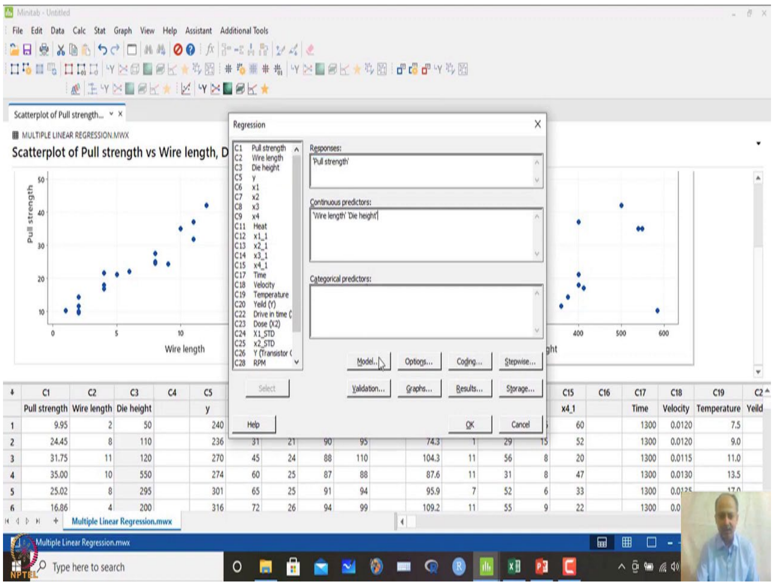
(Refer Slide Time: 20:55)



So, one of the diagrams is with wire length what you see is that, this is very strong relationship that we observe wire length with the strength and positive relationship exists. So, we expect the coefficient of beta should be positive and here it is somewhat not so prominent, but increasing trend we can see over here.

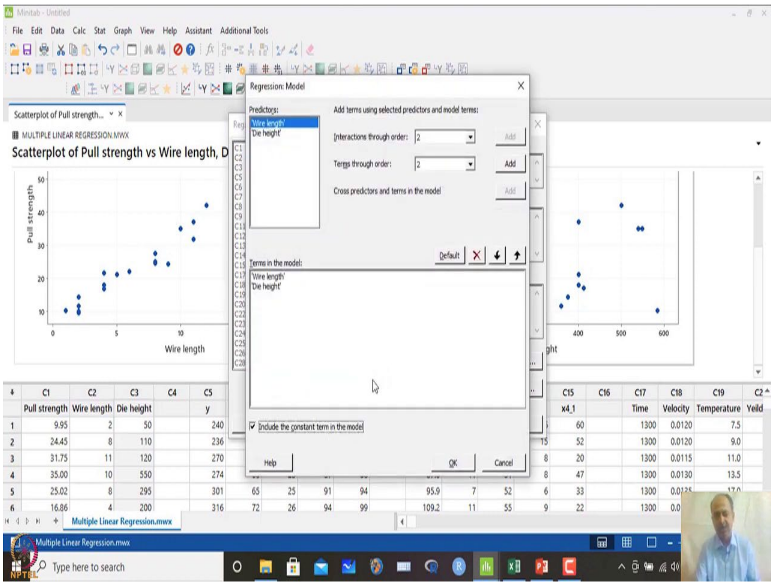
So, not so strong, but there is a positive relationship that we can understand and it exist and so, we should include that also in the model and after doing the scatter plot we understand, which is to be then we can go for regression.

(Refer Slide Time: 21:17)



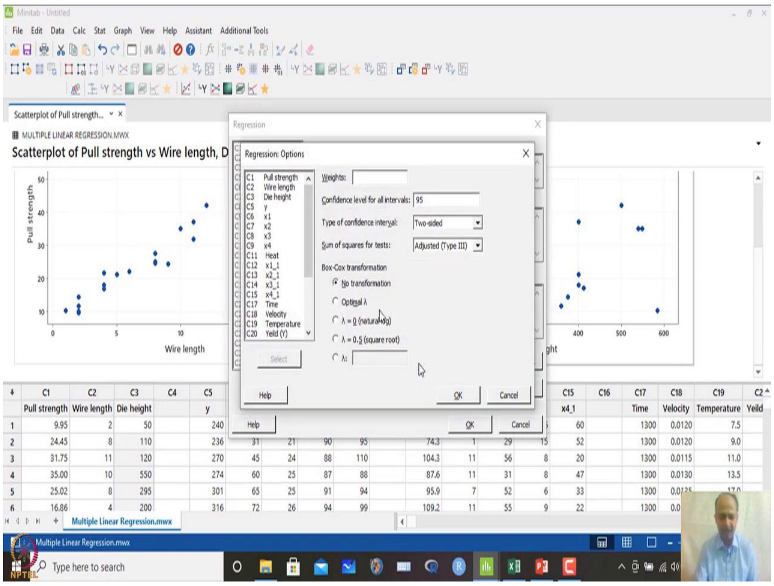
And fit regression and fit regression model what we will do is that, we will take pull strength as the response variable and continuous predictor it is continuous variable die height and wire length is placed over here.

(Refer Slide Time: 21:26)



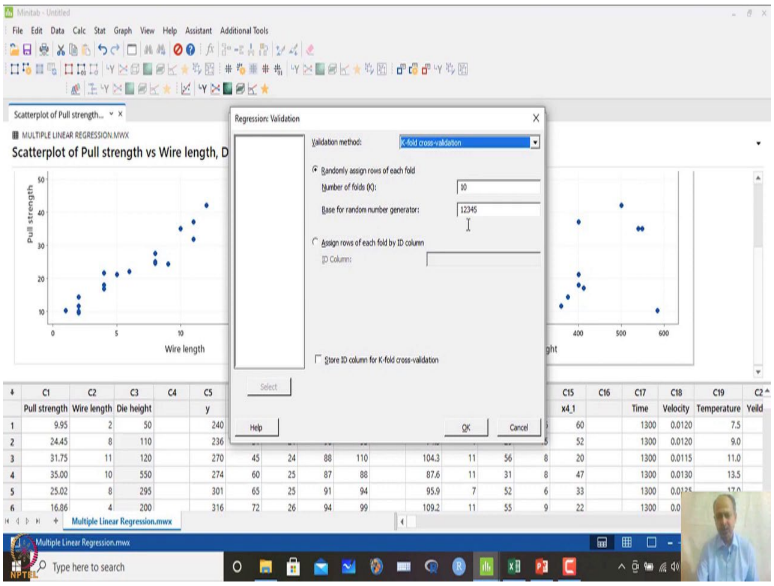
And in models what we will do is that we will include constant term that means, beta it has to be included that we have considered other things we are not changing over here. So, terms so we can add, but we you are not adding terms over here. So, we want linear equation at this present moment not polynomial at this time point.

(Refer Slide Time: 21:42)

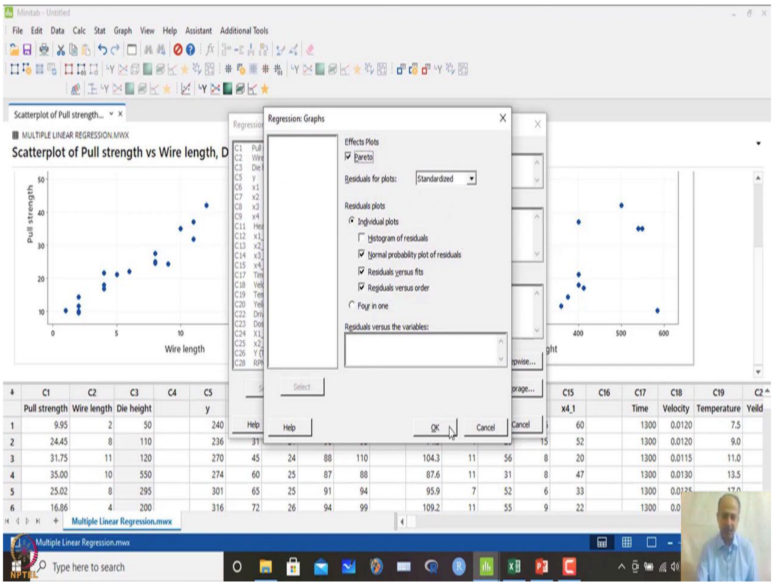


There if transformation is required we will see, if it is not required we do not need transformation. So, no transformation option is given ok. And coding we can avoid this time because; sometimes x variables are coded like that so, that helps in there is a theoretical advantage if we are doing coding over here ok, in design of experiment coding is done. So, that is an important aspect and step-wise regression we will try to understand afterward.

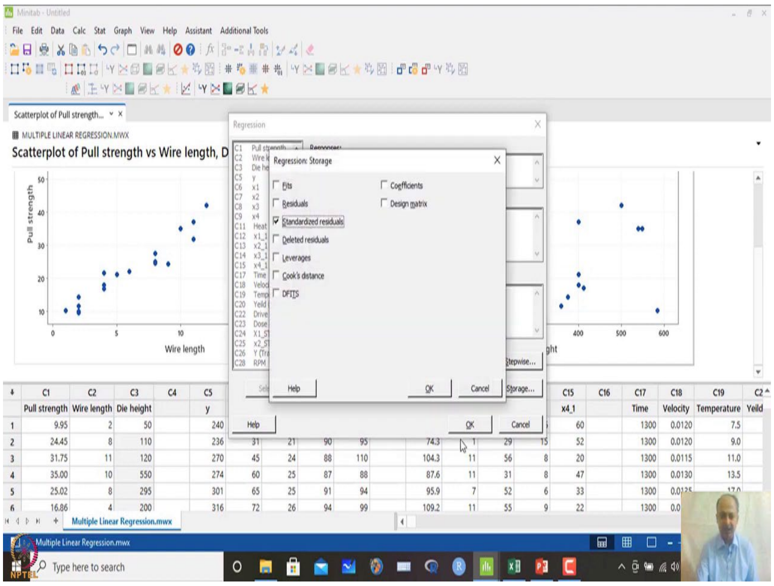
(Refer Slide Time: 22:10)



(Refer Slide Time: 22:17)

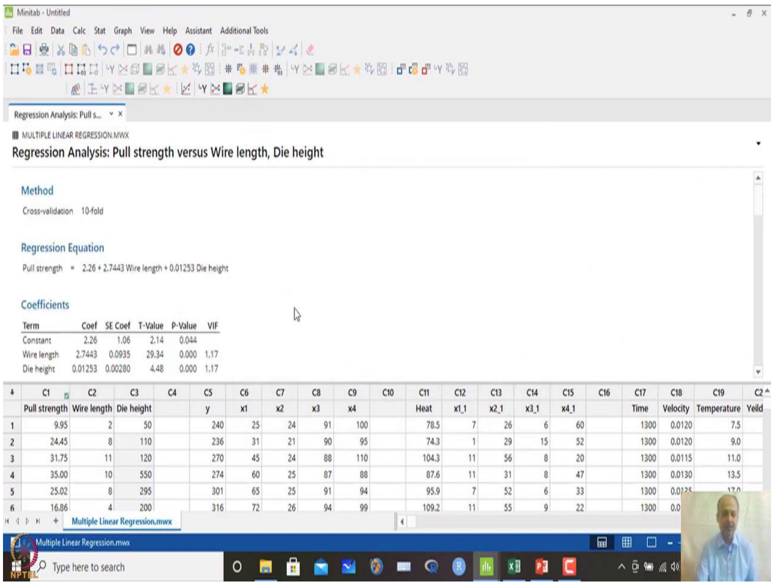


(Refer Slide Time: 22:27)

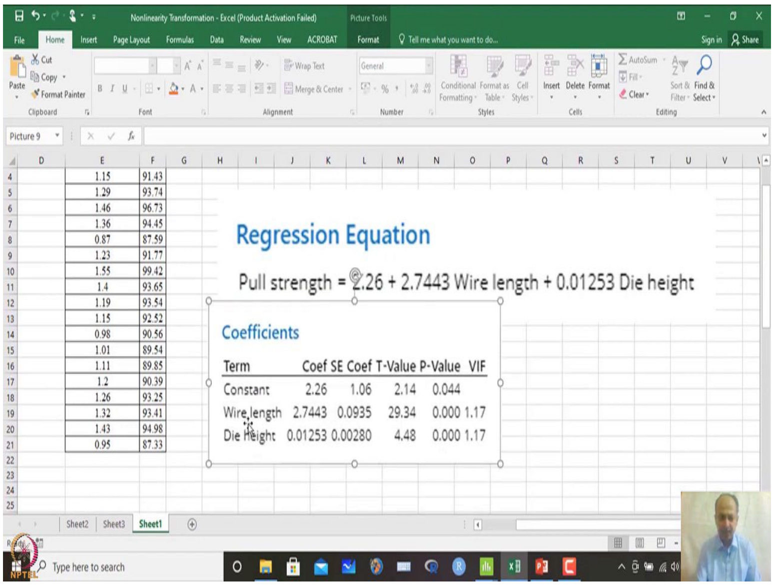


And model cross-validation is possible here also. I can go for K fold cross-validation, 10 is the K that we can assume over here, and then graphically we can see is normal plot residual versus fit residual versus order standardized residual and pareto effects whether, all the significant or not I will click ok and then, I can store the residual standardized residual over here and I click ok and then click ok over here and let us see what we are observing.

(Refer Slide Time: 22:33)



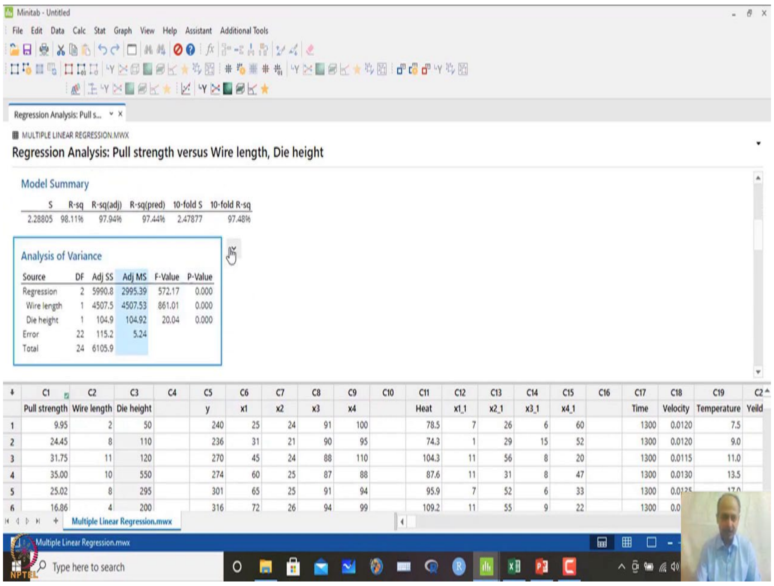
(Refer Slide Time: 22:44)



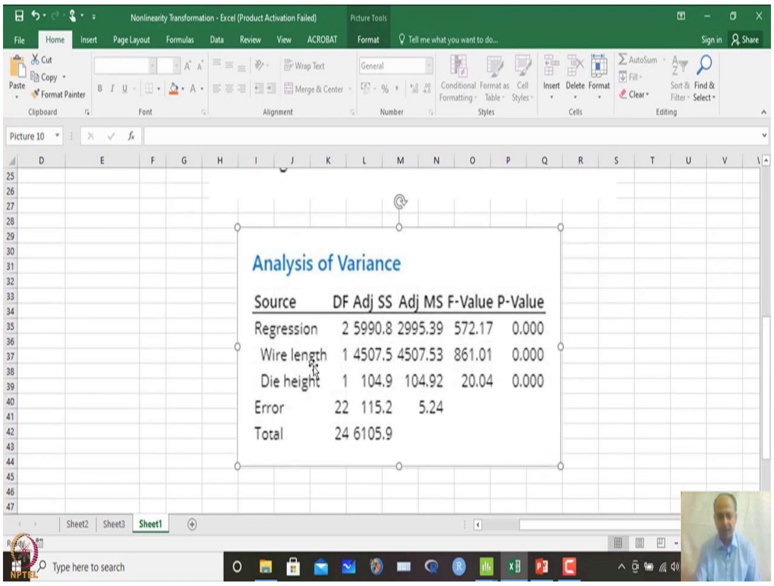
So, first is equation that is generated over here, MINITAB displays the equation and so, the equation will be given and we can just paste this over here and we can see that this equation is coming over here. So, this is the equation that we are having and then, next observation our observation is this is a coefficient information that we are getting.

So, this is the equation, second is model adequacy we want to see whether both the variables are important what we observe is that wire length is having a P-value, which is significant and also die height is having a P-value which is significant both the variables are important then, what we can do is that we can go to the ANOVA analysis and try to see.

(Refer Slide Time: 23:23)



(Refer Slide Time: 23:32)

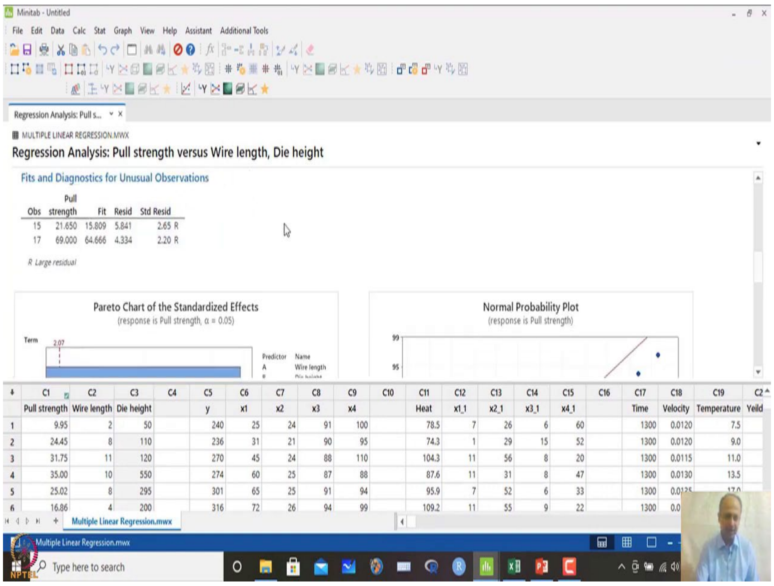


And so, this is the ANOVA analysis which we can copy and try to see this as an image and enlarge that and here also it will be if the coefficients are significant ANOVA will also show that one. So, wire length is coming out to be high value also F over here, die height is also high.

So, the P-value is significant that we are observing over here and both the variables are important. So, overall regression there is significant so, over all regression indicates that any one of the variable is significant and that that shows over here in the regression and individually if you have to see each of this P-value we can observe and we can see both are significant over here.

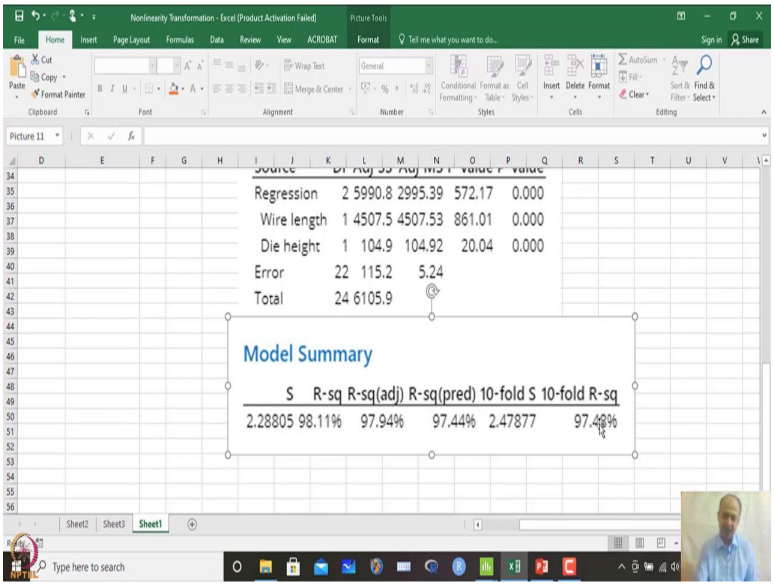
And degrees of freedom regressions what you see over here is that, there are 2 variables so, we have 2 degrees of freedom for these two x variables so, 2 degrees of freedom and each is consuming 1 degree of freedom over here, total number of observations is total degree of freedom 25th or 25 observations we have so, (25-1). And error is the remaining degree of freedom that we can calculate ok.

(Refer Slide Time: 24:26)



And then, it indicates that both are important for us and there are some unusual observation that you there you can see and K fold cross-validation, if you want to see this one copy as image and you can paste this one.

(Refer Slide Time: 24:37)

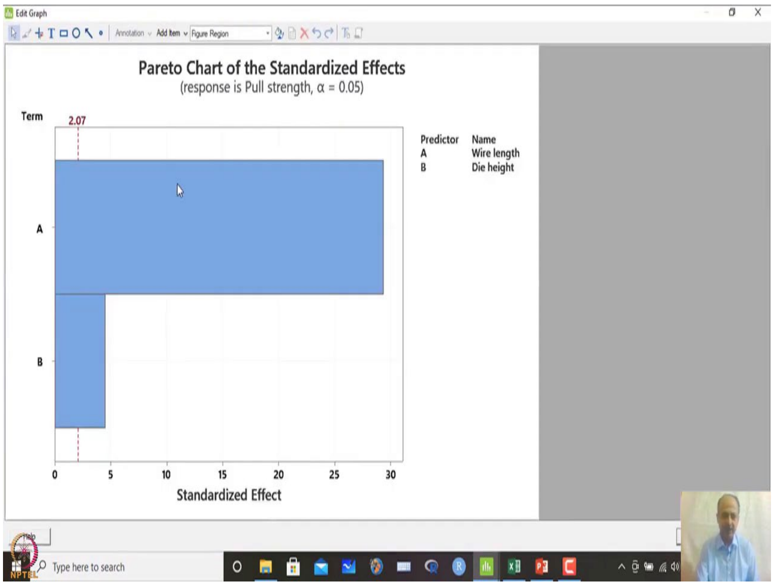


And you can see that, after cross-validation also model giving good results. So, overall R square adjusted is around 97 that is very good, we will see R square adjusted value over here I told. R square predicted should be closed close to R square adjusted, if the model is correct and can be generalized and 10-fold cross-validation and R square values of that.

Cross-validation means, we are taking one 10, we have divided the data in 10-folds 10 different data sets randomly and one of the dataset is used as testing and the other data set will generate the training data set which is the training data set and based on that we will generate the equation and then, test what is the R square values for the test data set like that.

So, that will be calculated for this and then what is reported is 10-fold R square values that is 97 that is quite good that is quite good over here, what we are observing.

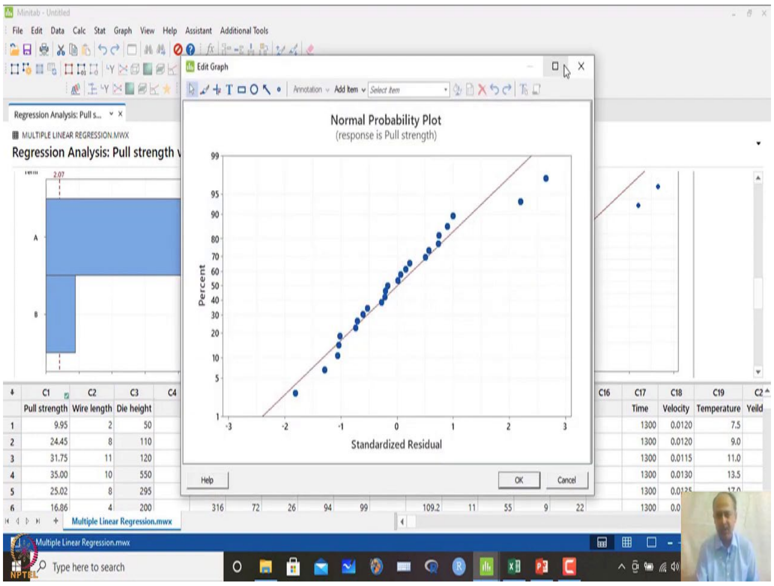
(Refer Slide Time: 25:25)



And then what we see is that, we have the plots also and plots if you see the effect plots what you see is that this at level of alpha 0.05 and this is the cut off over here that you see; and this calculations you can just check formula to do this calculation 2.07 how it and this is given in MINITAB.

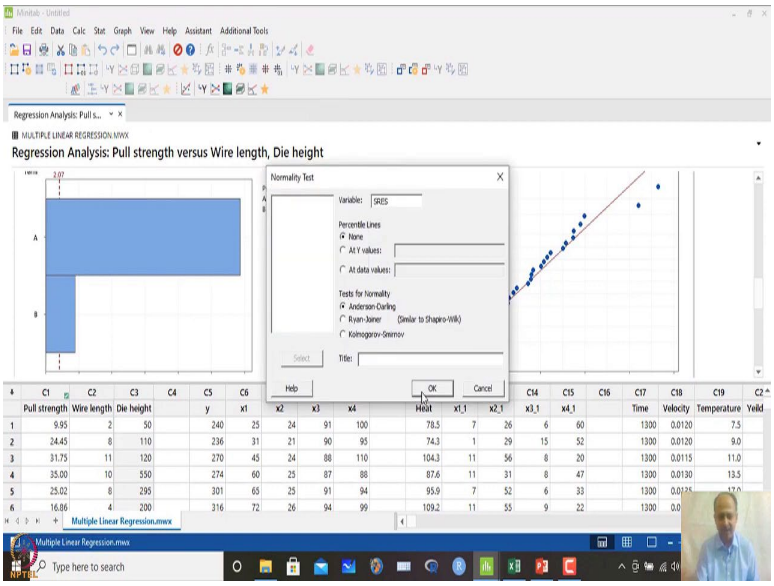
So in this case, and also any general books on how they are how they are calculating these values of cut off over here. So, A is significant which is beyond this cut-off, B is also significant which is beyond is both are important and can be included in the regression equation.

(Refer Slide Time: 25:55)

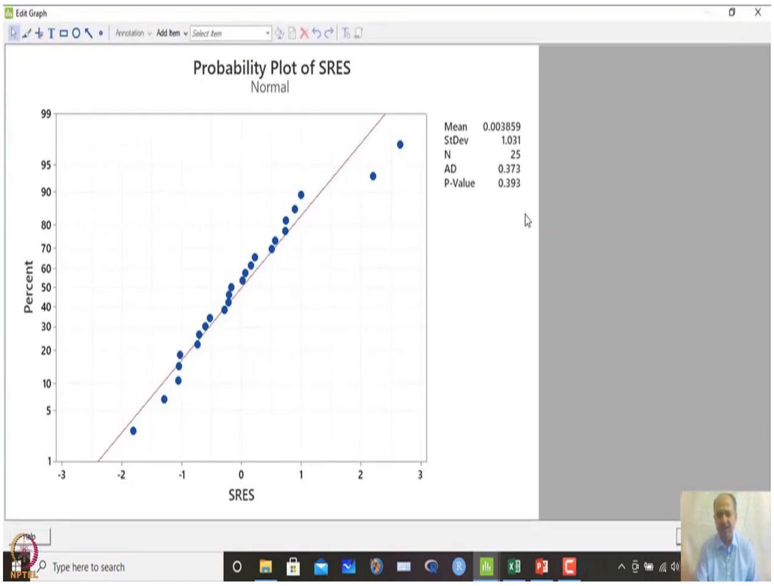


Also normal probability plot you see not much deviation, which most of the observations are middle part.

(Refer Slide Time: 26:08)



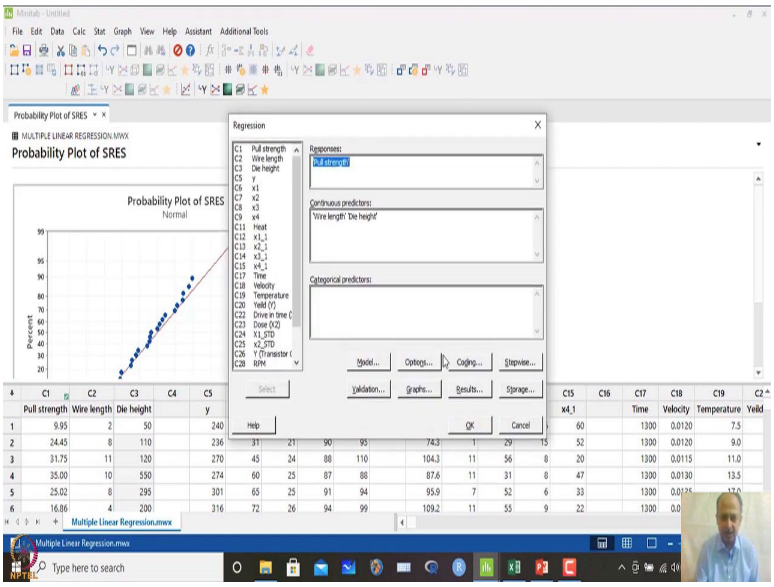
(Refer Slide Time: 26:11)



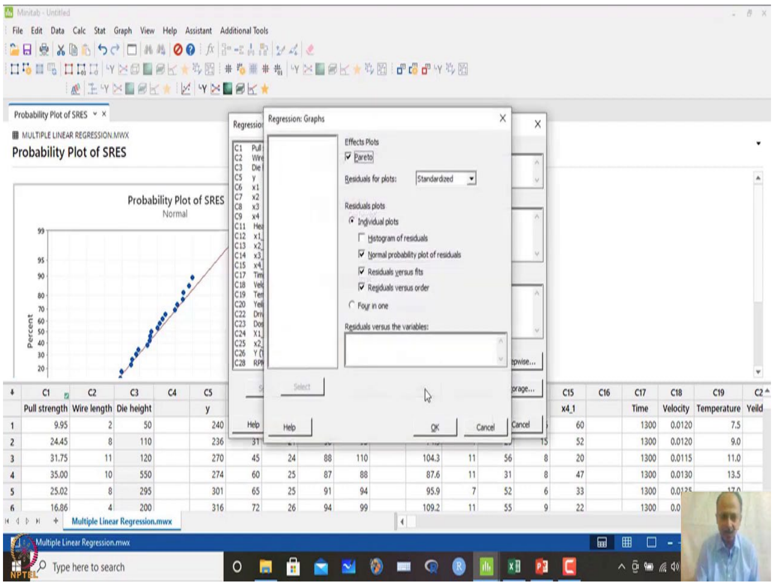
So, there is not much deviation, but we have save the residual and we can just check whether the basic residuals follow normality assumptions at the end you will find the residual and this is saved already and you go to Anderson-Darling test and Anderson-Darling test on this dataset indicates that the P-value is not significant.

So, in this case the error residuals is more or less, we can assume that that is ok. And what we can do is that we have not generated the heteroscedastic plots. So, in graph what we can do is that residual versus order residual versus order.

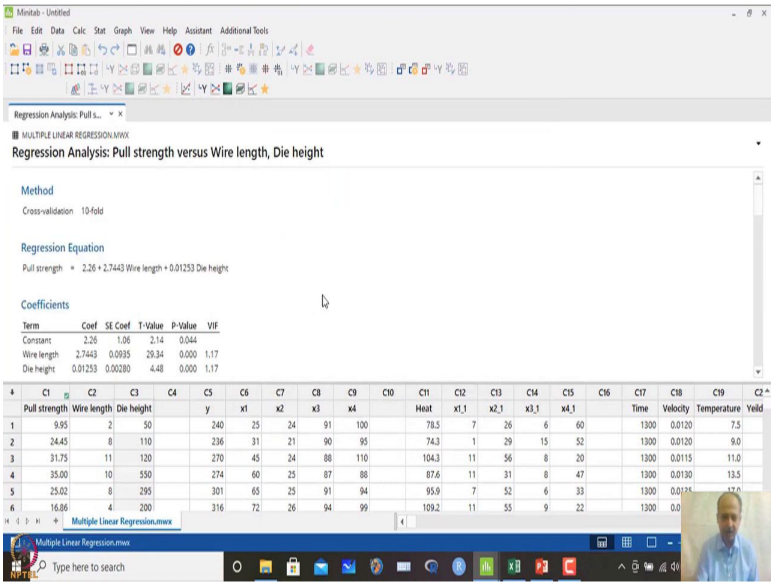
(Refer Slide Time: 26:33)



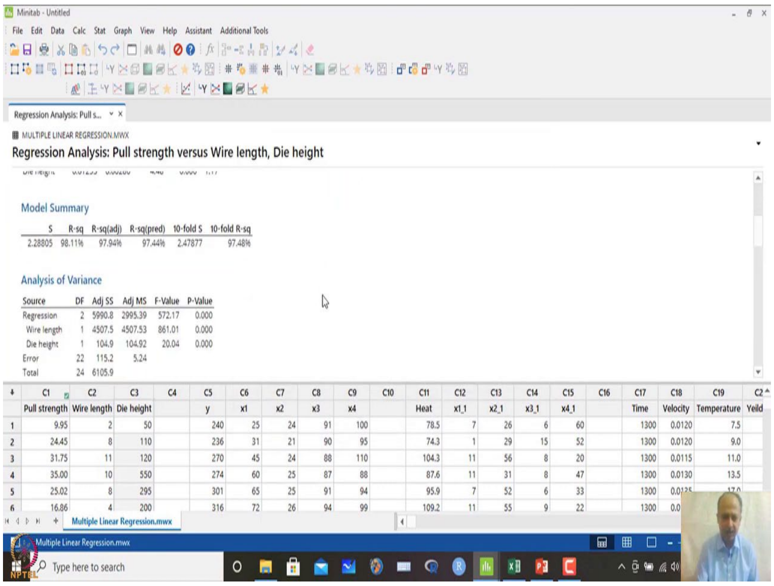
(Refer Slide Time: 26:35)



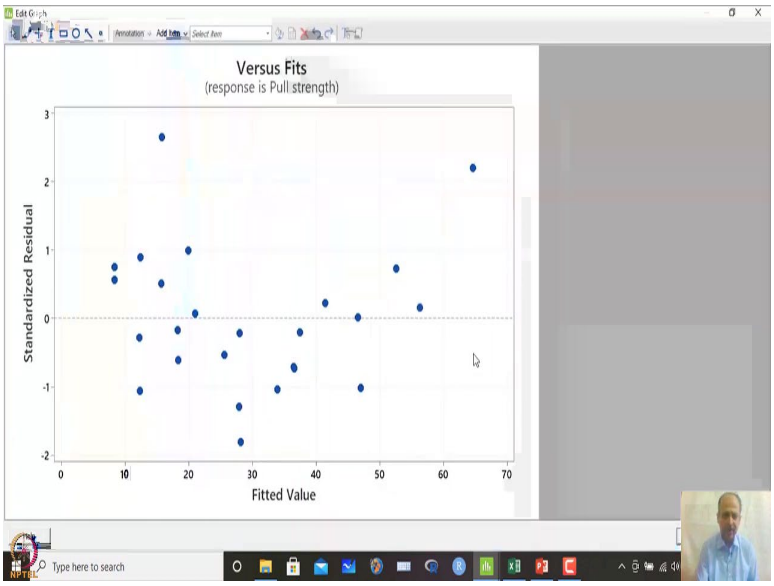
(Refer Slide Time: 26:40)



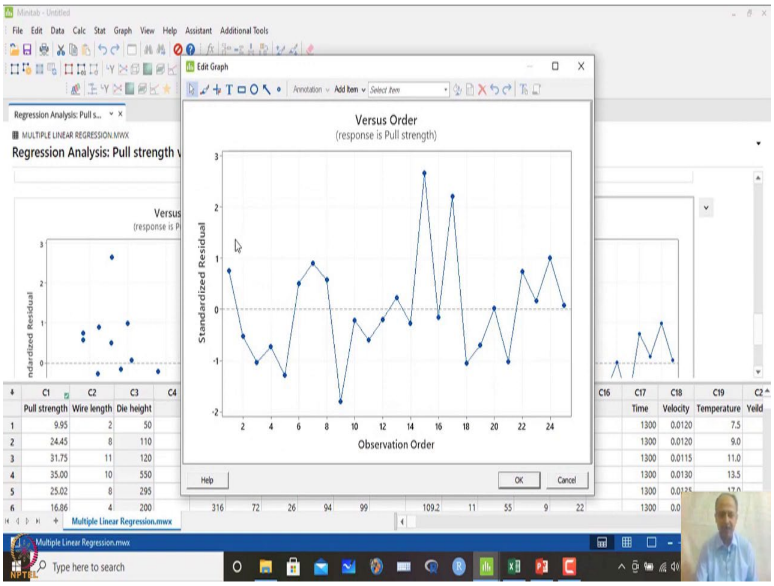
(Refer Slide Time: 26:40)



(Refer Slide Time: 26:41)

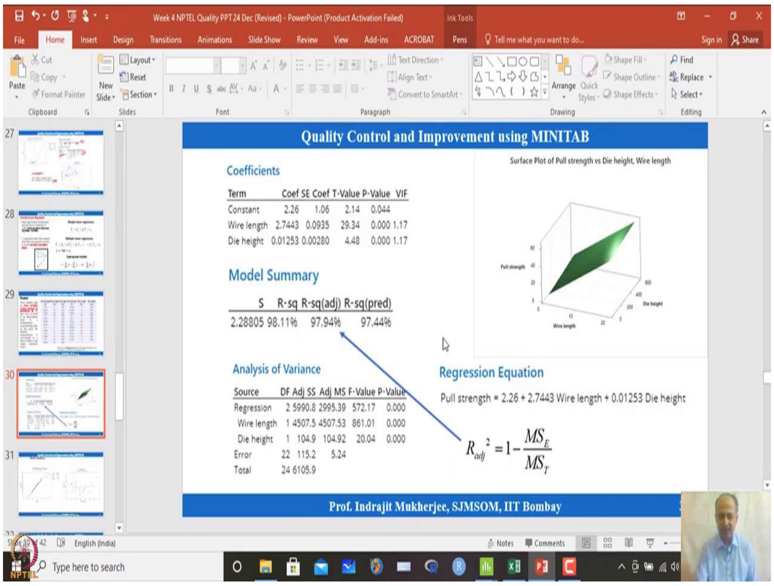


(Refer Slide Time: 26:57)



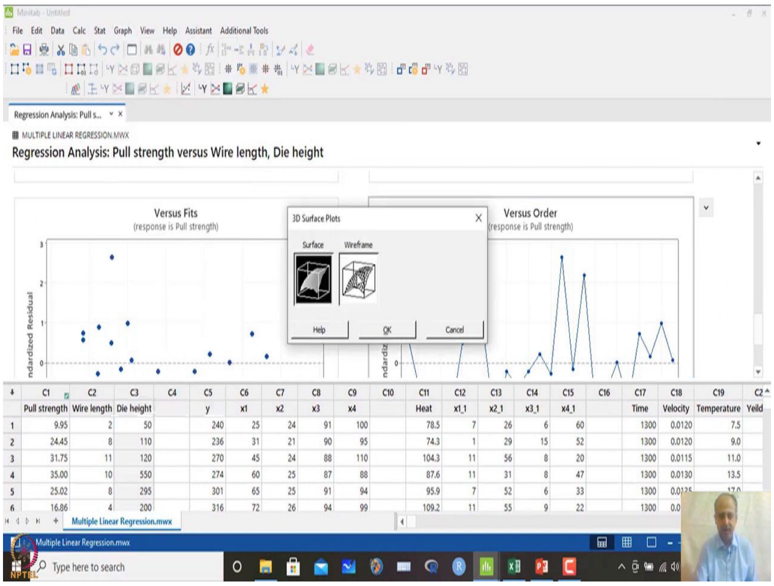
So, we have not done it. So, we can just check that on graphically. So, this is the graph that we are generating based on this data set, not much unusual observation. So, Breusch Pagan test can be done over here and try to confirm whether there is any heteroscedasticity behavior of the residuals that can be observed, but I do not think it is there and also there is no as such strength that we observe in the standardized residual with observed order over here.

(Refer Slide Time: 27:06)

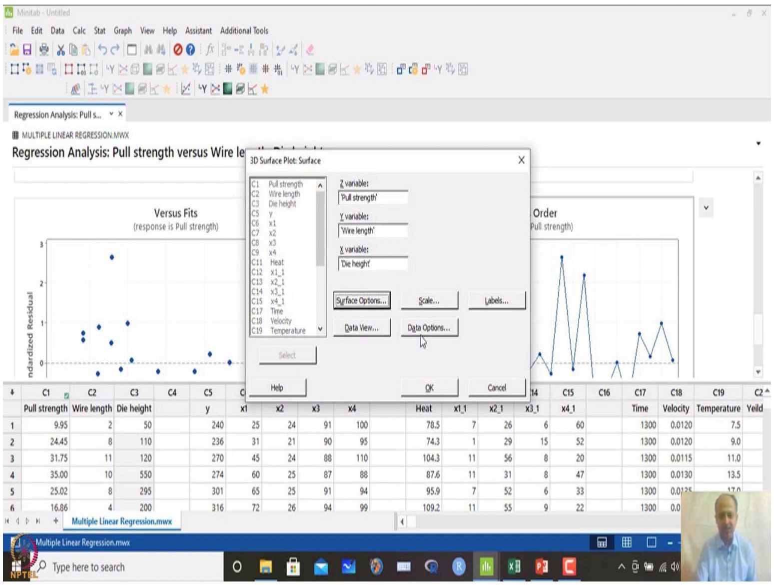


So, in this case also not so prominent, but we can do that individual testing like that. So, in this case and also we can do the surface plot that I have not shown.

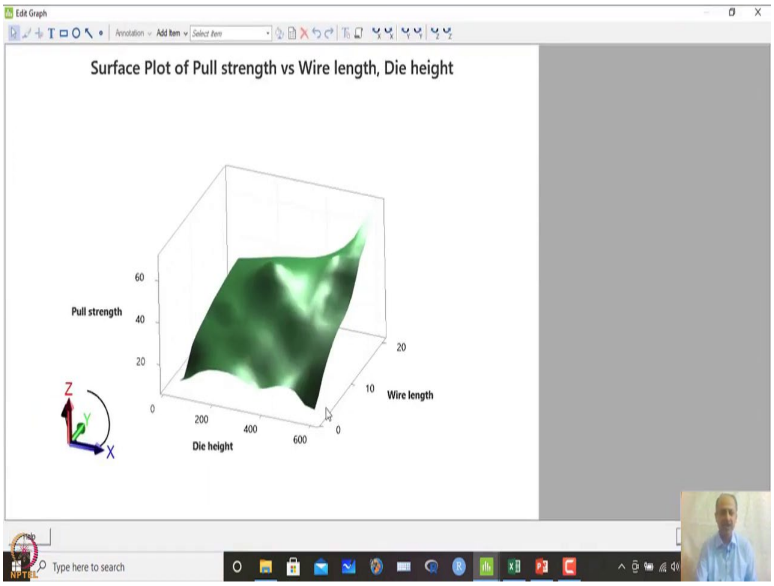
(Refer Slide Time: 27:16)



(Refer Slide Time: 27:21)



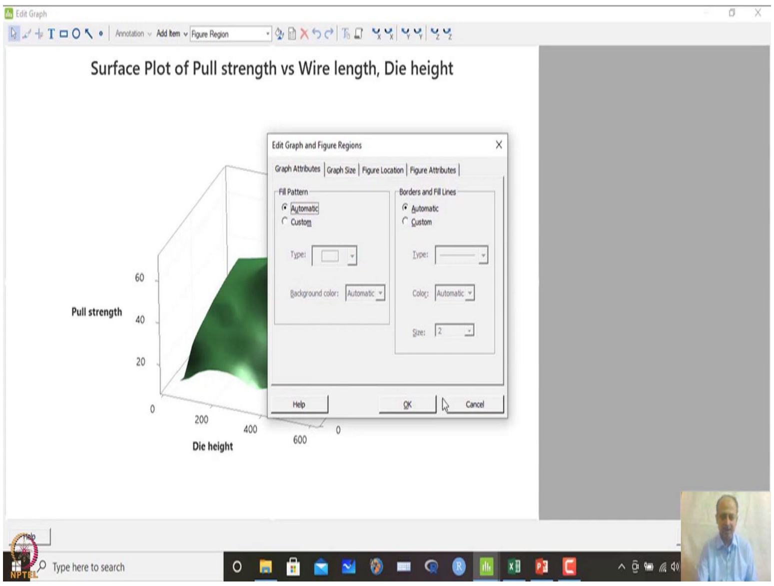
(Refer Slide Time: 27:29)



So, surface plot can also be done, graph over here 3D scatterplot 3D surface plot is option is there. So, I can do a wire diagram or I can do surface plot over here, if you click ok it will tell which is the Z direction which is a variable. So, this is x 1 and x2 variables like that, and you can change the shapes you can change the plots types of plots like this.

So, this is the plot that will be generated surface plot and you can just rotate this one also. So, you can rotate the axis like that. So, that is also possible over here.

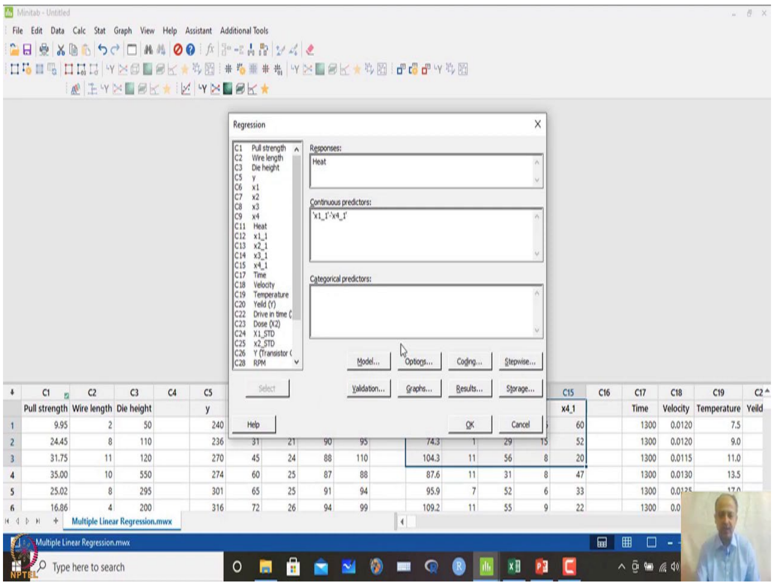
(Refer Slide Time: 27:38)



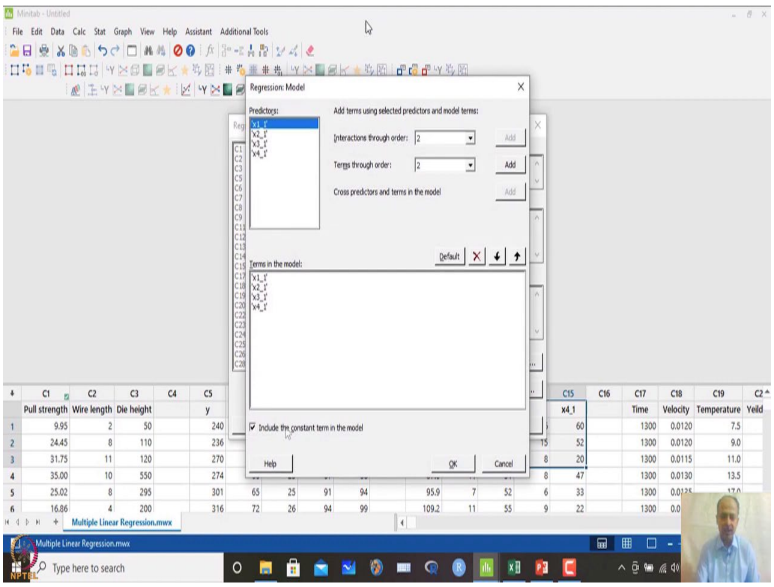
So, I can rotate that is possible over here. So, this is if you want to rotate and see what is happening over here. So, this is also possible in MINITAB and that gives you an idea, how the surface is over here ok.

So, this is also possible in MINITAB this is also possible in MINITAB. So, that we can see all aspects of the regressions all aspects of the regression over here. Similarly there is another data set like this heat and with some variable x 1 to x 4 and we want to see the relationship between this.

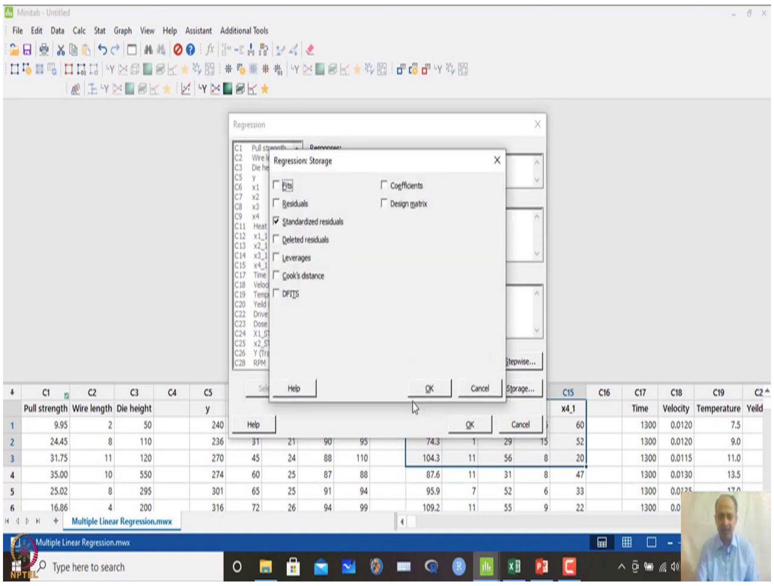
(Refer Slide Time: 28:17)



(Refer Slide Time: 28:33)



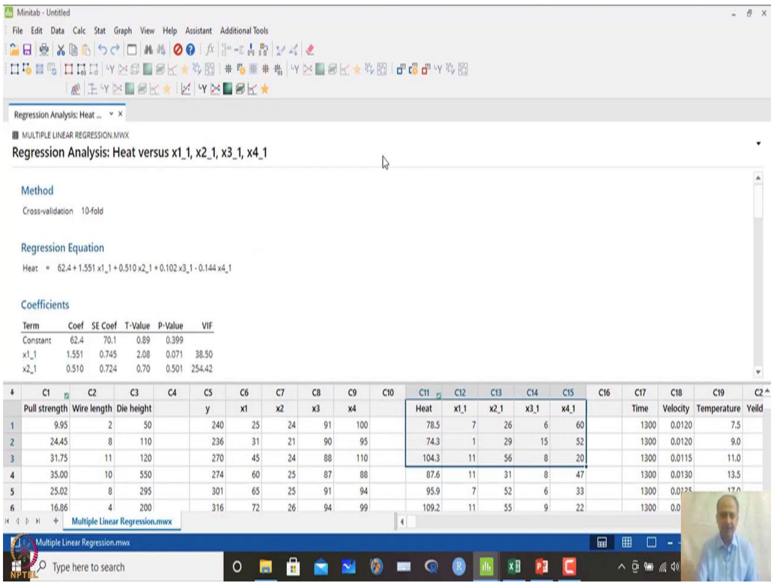
(Refer Slide Time: 28:49)



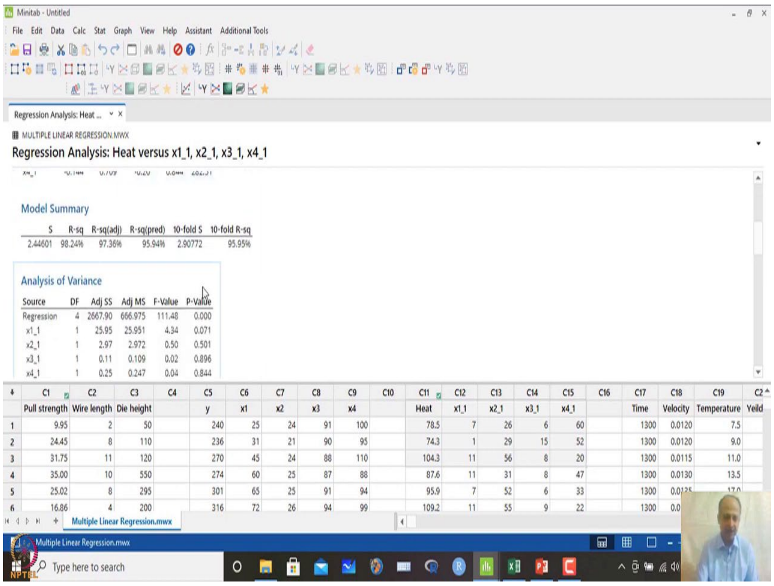
So, we can close this one and we can see another examples of heat and this. So, if you want to generate this one regression equation regression fit regression. So, in this case variables will only change so, heat is the outcome that is measured over here and these are the x1 to x4 variables that you want to include in the model select that one. And then in model include the constant terms that is there and we do not want to increase any other terms over here.

So, this is ok and in options no transformation because everything is let us assume that everything is ok. So, then validation and everything store residuals over here. So, you can store the residuals and you click ok finally, when you click ok.

(Refer Slide Time: 28:52)



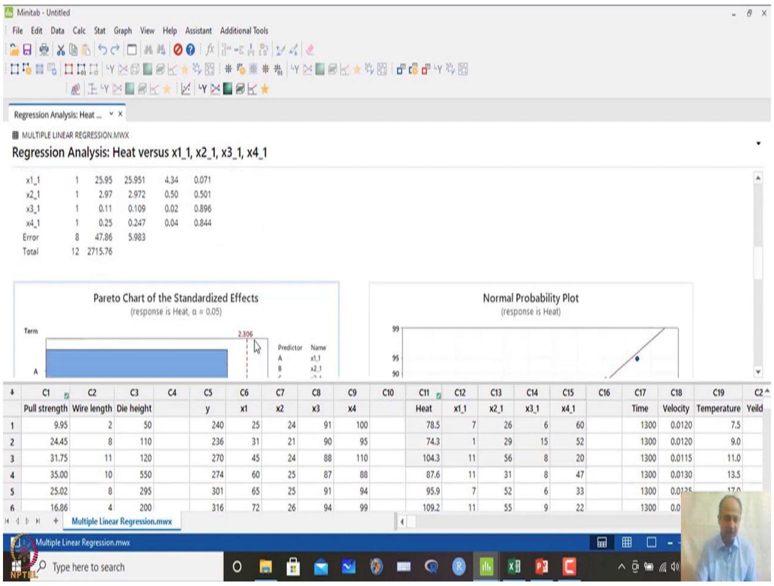
(Refer Slide Time: 28:55)



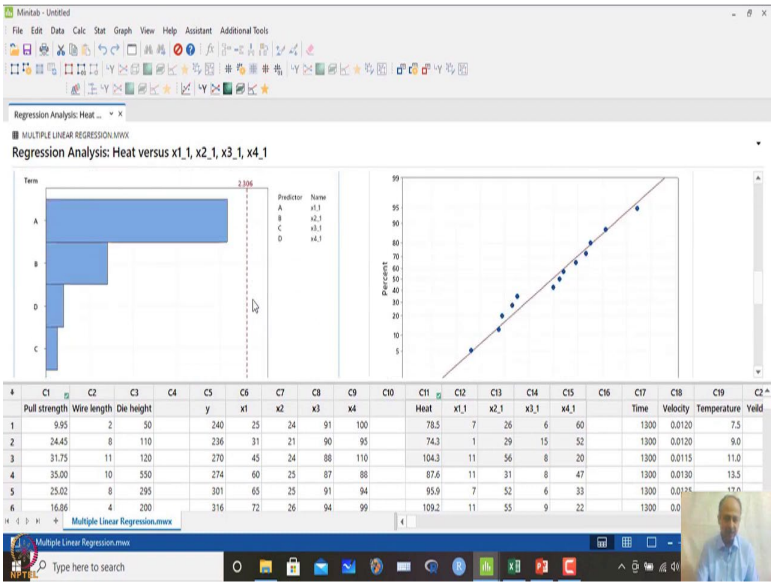
So, then what happens? The equation is given at the first stage coefficient, which is significant which is not over here. So, you see x1 is significant over x1 is also not significant none of the variable is coming out to be significant over here. So in this case, what is happening is that all the P-values, but in this case ok.

So, but there is a clear relationship R square adjusted value is 97.36 and K fold cross-validation is also high ok. So, in this scenario what we are getting is that x1 is quite prominent, x1 is values is quite prominent over here.

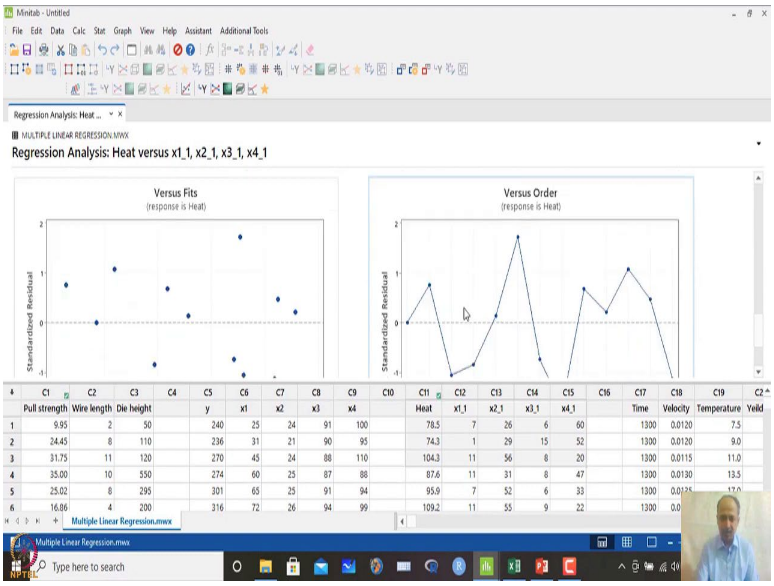
(Refer Slide Time: 29:27)



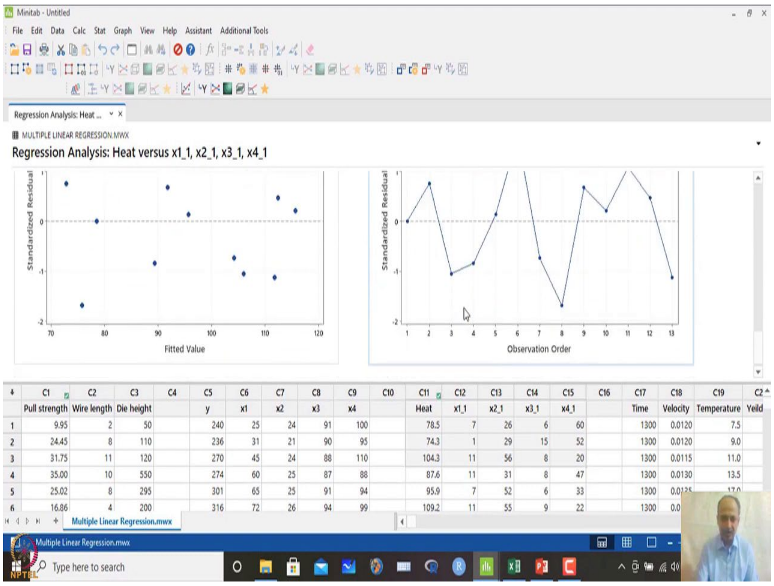
(Refer Slide Time: 29:31)



(Refer Slide Time: 29:32)

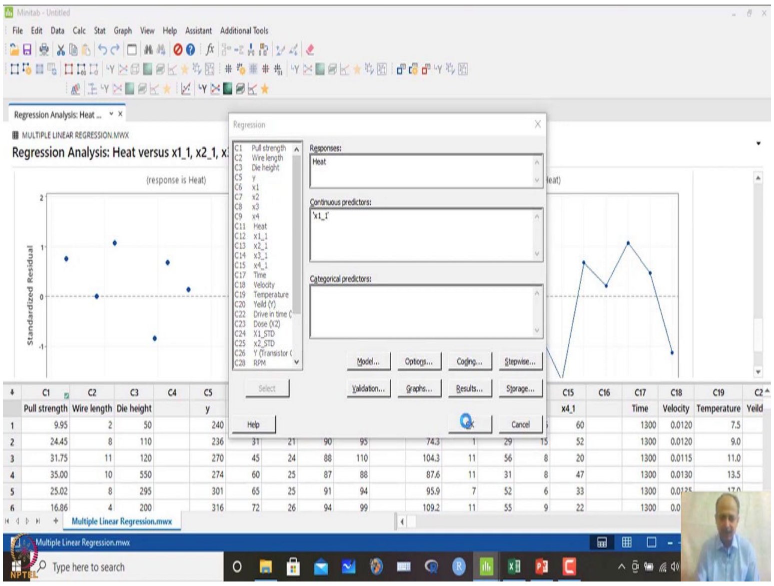


(Refer Slide Time: 29:33)

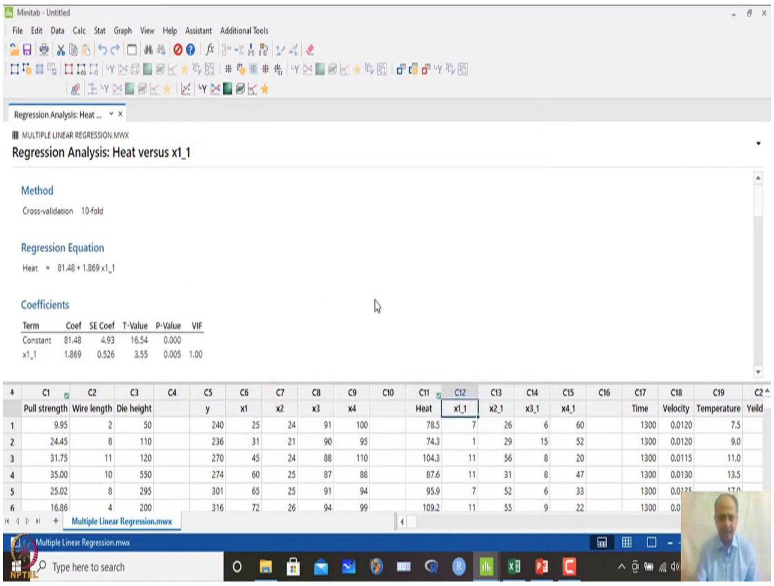


And the variables what we are seeing is that, there is some issues over here in this case and there is no issues other than that. So, if we have included only one variable over here x1 what happens, that we can see x1 over here only x1, if we consider only x1 over here, what is the scenario?

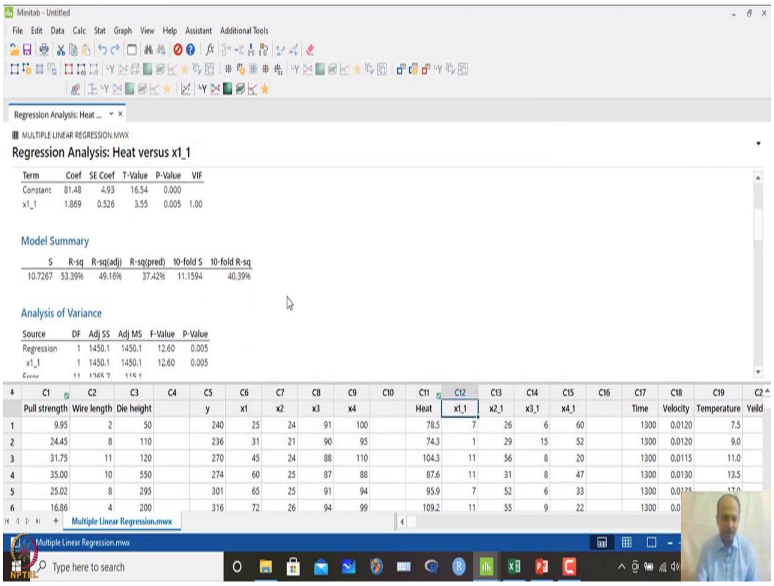
(Refer Slide Time: 29:49)



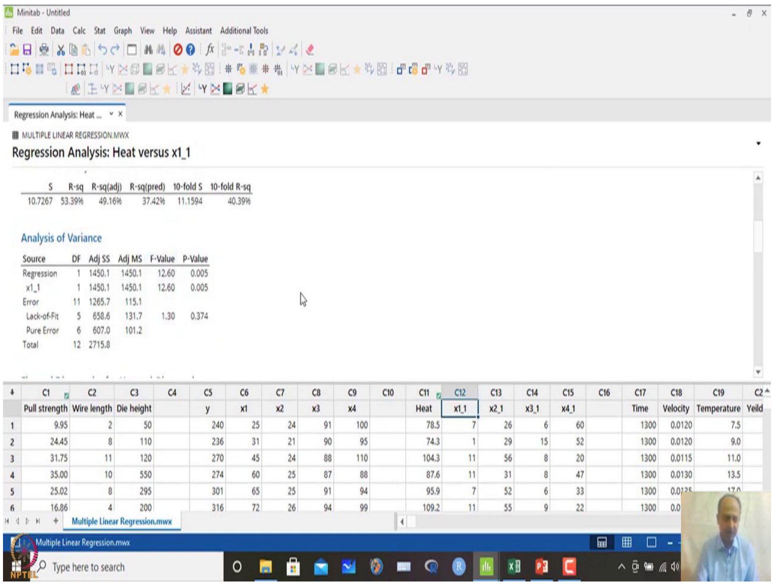
(Refer Slide Time: 29:54)



(Refer Slide Time: 29:55)



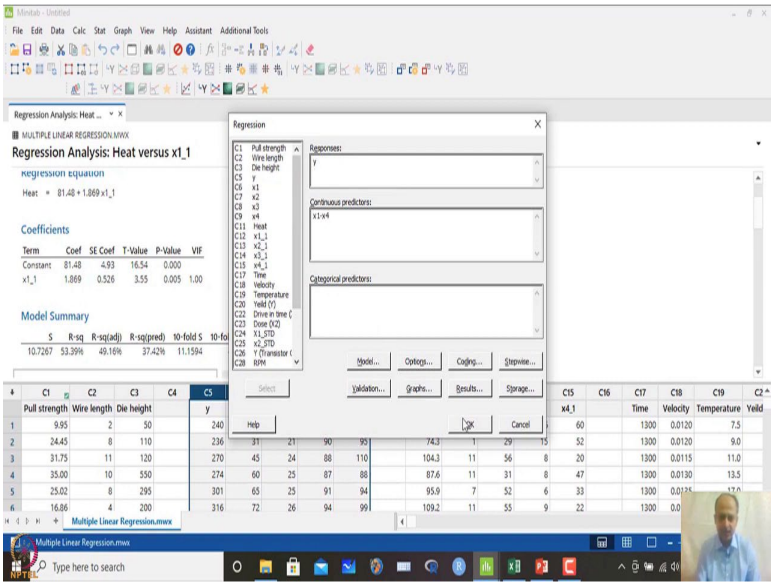
(Refer Slide Time: 29:58)



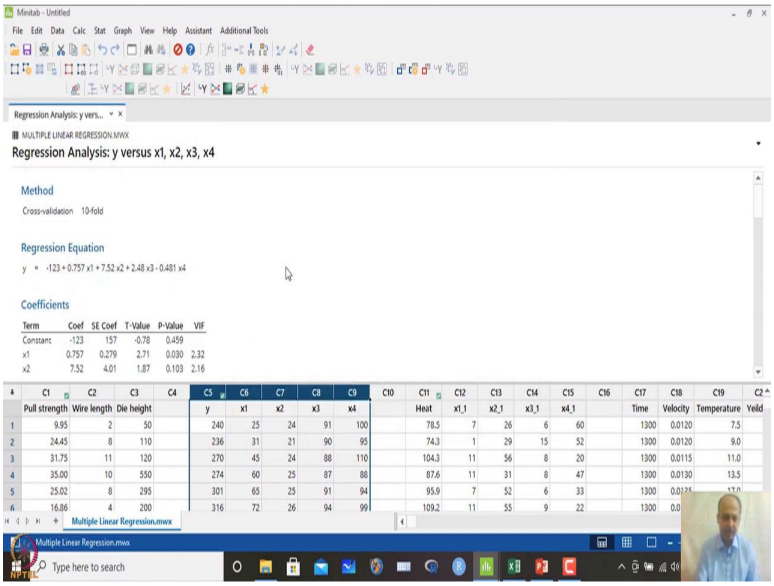
Fit regression in case of all variables, we can only consider x1 variables over here. And I click ok, and what I will observe is that not much variable it is explain R square adjusted is low. So, we keep on adding the variables and there is no lack of fit also in this, but overall explanation of the total variability. So, we keep on adding the variables x1 is added, that explains some part of the variability when I add x2 that add some part variability x3 and x4 like that.

So, this kind of analysis simple analysis can be done let us also assume these C5 to C9 variables over here, and in this case also we can see the regression equation how we which what is the equation that is generates over here.

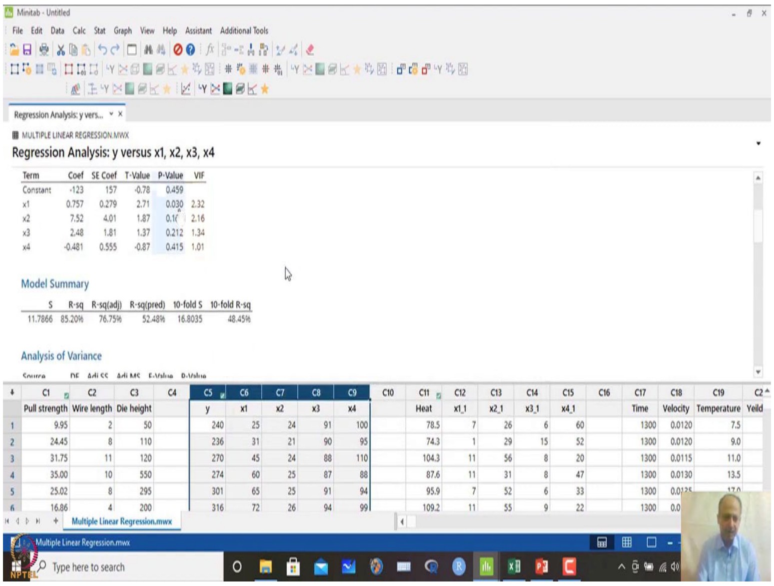
(Refer Slide Time: 30:37)



(Refer Slide Time: 30:50)



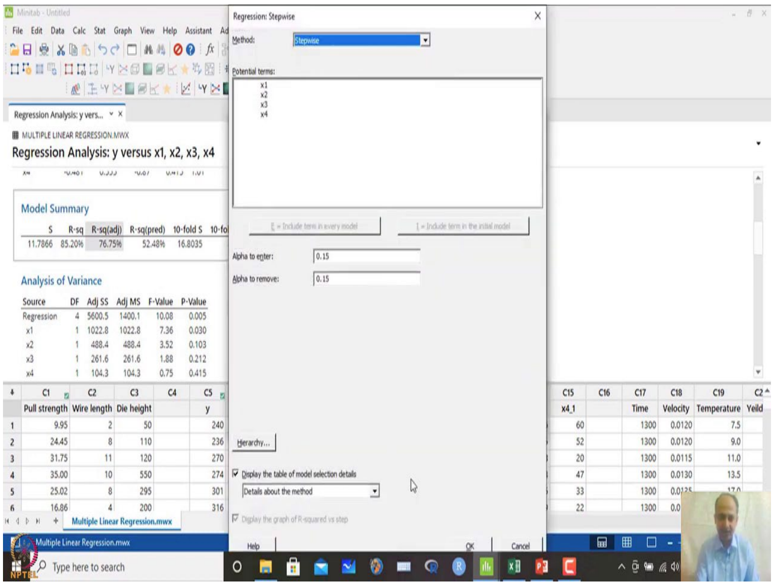
(Refer Slide Time: 30:50)



So, this what we can do is that, we can just go for this Y variable that is over here and these are the x1 to x4 over here and this is selected sorry this is selected over here in the continuous predictor x1 to x4. So, this is selected over here and if I go for this what happens is that, here only one variable comes out x1 to be prominent over here P-value is not significant over here, but we are retaining those variables because that gives me R square adjusted value.

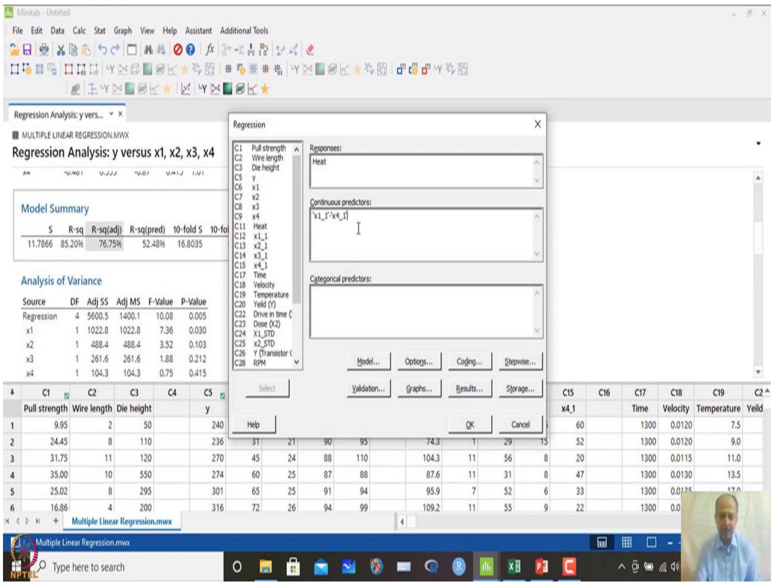
And we can do it automatically also. So, let us try to also discuss this one, how do we select which is the variable to be included which is excluded. So, regression analysis also gives you a fit regression model.

(Refer Slide Time: 31:14)

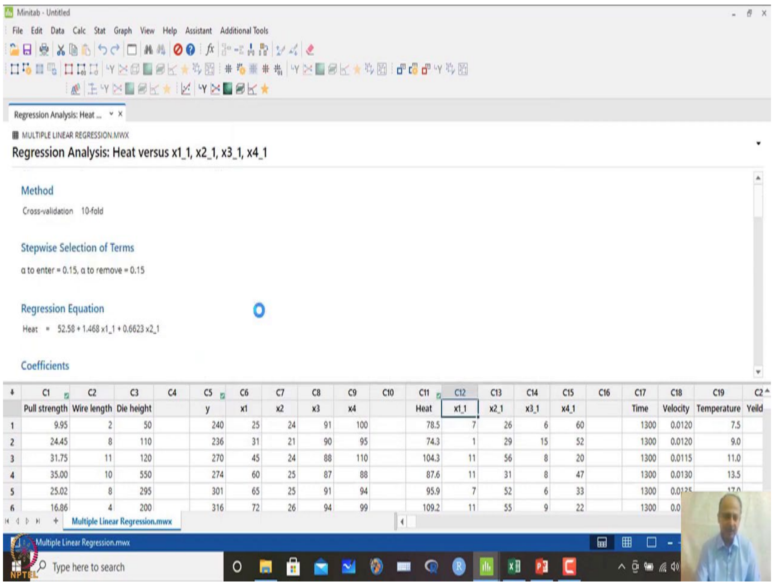


There is a option of step-wise regression over here. So, if I go for step-wise regression over here, it will automatically suggest which variables to keep and which variables not to keep like that.

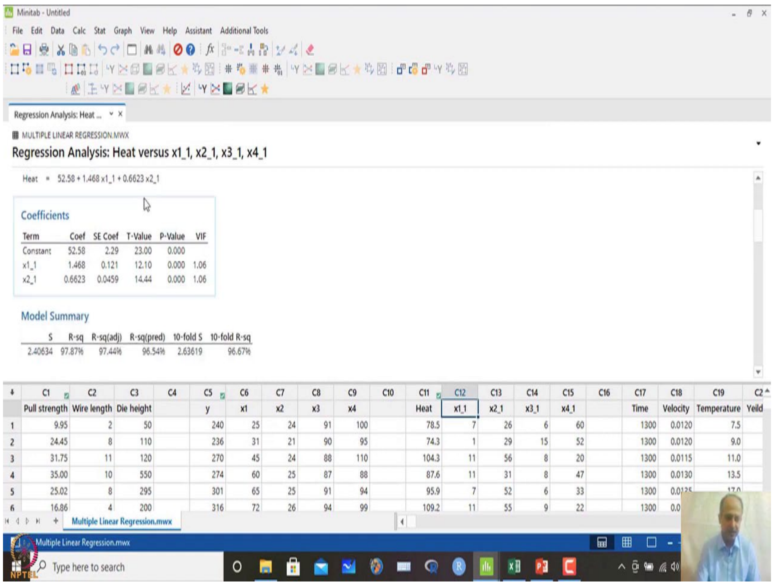
(Refer Slide Time: 31:23)



(Refer Slide Time: 31:33)



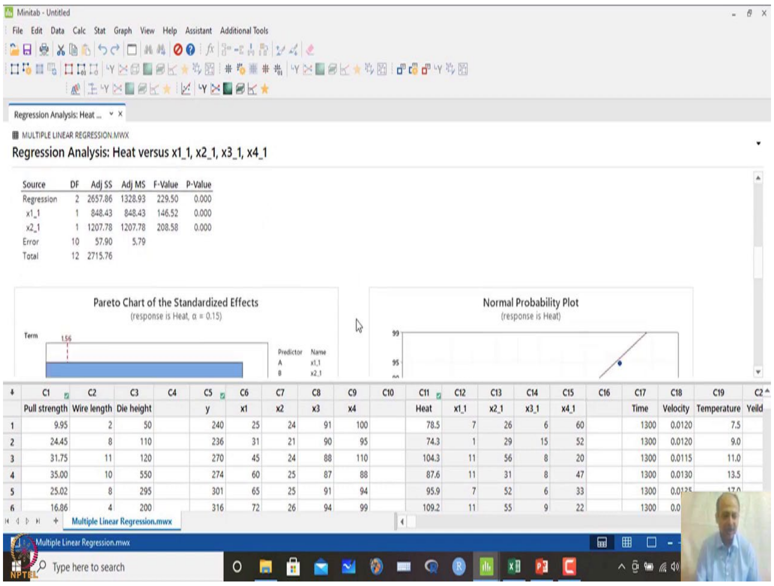
(Refer Slide Time: 31:37)



So, if I change these variables over here let us say heat is the first variable and we are going for x1 to x4 over here, and I click ok and in this case if I click ok it will suggest which of the variables to be taken and you see that it has only considered one variable that is first variable x sorry second variable it is considered to be placed over here.

So, there is certainly some other problems that is existing and x1 and x2. So, x1 and xs2 is prominent. So, these two variables are included over here. So, these are the two variables to be included, we do not want to include all variables.

(Refer Slide Time: 31:55)

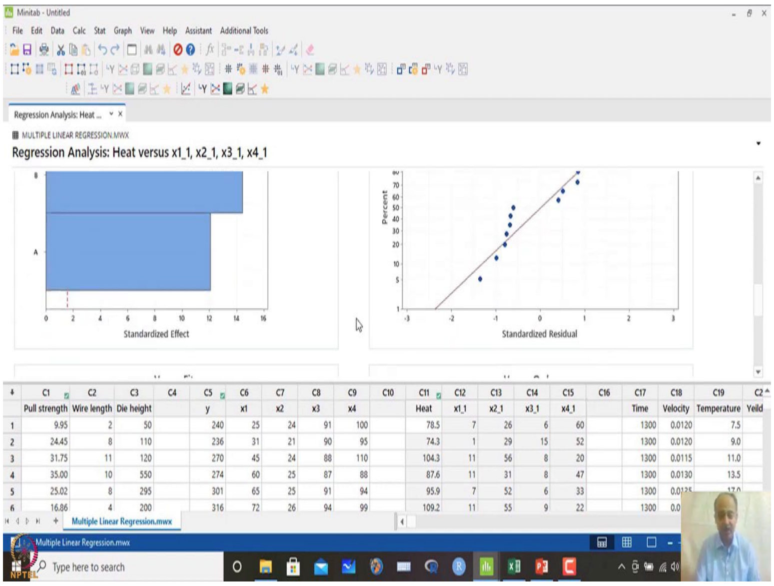


So, we will discuss more about this stepwise regression and you see whenever there is a like previously what we were facing is that which variable we should take because they are not prominent. So, which one will go in which one will go out like that we are not certain like that.

So, what we do is that, we have a method which is known as stepwise regression and best subset regression, which will allow us to identify which variables to include in the model, which variables to in exclude in the model. Because, I was not sure that it is not significant all are not significant.

So, in this case which one will go in and which will give me the best fit like that. So, MINITAB this technique which is known as stepwise automatically identifies, which two variable basically maximizes the R square predicted and R square adjusted value over here. So, this is around 97 percent.

(Refer Slide Time: 32:41)



And this is the best model that you can get out of this dataset that you have generated. So, include x1 this one column C12 and C13 and that gives me an option that, which variables should go in and which variable which should not be included. So, we will continue with this discussion of best subset regression in scenarios when we are in dilemma which will go in and which will go out.

And then, extend that one and discuss about one topic which is known as multi co linearity in regression and that is a one I wanted to discuss. Because, that is that would significantly affects the model generalization basically. So, we will discuss about that in your next class ok.

So, we will start from here where we left we will take some example and how to include x variables like there should not be any dilemma this will go in this will go out, and no confusion in selecting of the variables. So, I have a easy way by going through the stepwise regression and using best subset regression like that, best subset methods to select the variables that will lead to the final generalized model like that.

So, we will stop here and we will continue in our next session.

Thank you.