

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

**Lecture - 24**  
**Linear Regressions**

Hello everyone and welcome to our session 24 on Quality Control and Improvement with MINITAB. So, I am Professor Indrajit Mukherjee from Shailesh J. Mehta School of Management IIT Bombay. So, in previous session what we are doing is that we are discussing on Basics of Regression and which is an important tool to identify variables which can be considered for further experimentation.

(Refer Slide Time: 00:46)

**Quality Control and Improvement using MINITAB**

**Simple Regression Model**

$y = f(x)$


Expected Response      Slope Coefficient      Independent Variable      Error in model estimation

$E(y|x) = \hat{\beta}_0 + \hat{\beta}_1(x) + \varepsilon$

Linear component      Random Error

Simple regression considers a single **regressor or a predictor** (x) and a **predicted or response variable** (y)

NPTEL      Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay



And so the basic models that regression uses is proposed by Gauss and what we have and this is the basic fundamental model that we are using over here to understand relationship between y and x. And here we are modeling expected value of y with respect to x over here. So, there are 2 coefficients that is similar in line with what we know about basic line equations that is  $y = mx + c$ , c is known as intercept of the models and m is known as slope of the models like that ok.

So, this is a simple linear regression. So, I am assuming one single x I am having and  $E(y|x)$  is the expectation of y for a given x this is the conditional values that is mean

values that we can expect ok. So, at different conditions of  $x$  what we will have? We have different values of  $y$  and the condition if I can change the condition and again reset that one to that same condition even the output will be defined.

So, that was observed in when we are talking about analysis of variance and so expected value of mean value is basically modeled over here with respect to  $x$ . So, this is the linear component that is  $y = mx + c$ , this is the component where  $m$  is we can think of a slope over here.

So, for 1 unit 1 unit change in  $x$  what is the change in  $y$  expected value of  $y$  that is basically  $\beta_1$  over here and if you just extrapolate this regression equation if I have developed something  $y$  is  $\beta_0$  and if you can extend the line what is the intercept? That is this is the  $\beta_0$  intercept over here. So, this is this will be the  $\beta_0$  component over here, the value of  $y$  expected value of  $y$  when  $x$  is equals to 0 basically.

So, this is the but generally in regression we do not; we do not extrapolate and we do not extend that one, but this is the intercept concept that we have over here. So, physical interpretation is not possible for  $\beta_0$ , but  $\beta_1$  has a physical interpretation like that for every unit increase in  $x$ . What is the expected value change in  $y$  basically change in  $y$ ? So, that is the interpretation.

So,  $\beta_0$  and  $\beta_1$  are the two important parameters which needs to be estimated from this model. So, if I can get the values then I can write the function over here and if I have the value of  $\beta_0$  and  $\beta_1$ .

(Refer Slide Time: 02:57)

### Quality Control and Improvement using MINITAB

The regression equation is only an **approximation** to the **true functional** relationship between the variables.

- Regression model: **Empirical model**

- Two important objectives:

- **Estimate the unknown parameters** (fitting the model to the data): The **method of least squares**.  
 $\hat{\beta}_0, \hat{\beta}_1$

- **Model adequacy checking**: An iterative procedure to choose an appropriate regression model to describe the data.



Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay



So, how  $\beta_0$  and  $\beta_1$  is estimated that is important for us and so estimation is important over here. So, estimation of unknown parameters that is  $\beta_0$  and  $\beta_1$  I need to estimate this one. So, I am writing  $\hat{\beta}_0$  hat and  $\hat{\beta}_1$  hat and whenever we have estimated that one then we have to do some model adequacy checks, that is also necessary like in nano work we have done.

Similar kind of tests are required in this regression analysis, I told that regression is an extension of analysis of variance basically. So, that is also important model adequacy checks like that ok.

(Refer Slide Time: 03:25)

**Quality Control and Improvement using MINITAB**

**Least Square Estimation**

$\beta_0$  and  $\beta_1$  are obtained by finding the values of  $\beta_0$  and  $\beta_1$  that **minimize the sum of the squared residuals**

$$\sum_{i=1}^n e_i^2 = \sum (\hat{y} - \hat{y})^2$$

$$= \sum [y - (\beta_0 + \beta_1 x)]^2$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \left( \sum_{i=1}^n y_i \right) \left( \sum_{i=1}^n x_i \right) / n}{\sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 / n}$$

Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay

So, this is the mathematical equation. So, I have a so what it will do is that many many lines can be constructed. So, it will plot the lines and innumerable line out of the innumerable lines. What it will do is that where the error is minimized basically, where the predicted value with the actual value so this will give you an error.

So, error minimization basically how can I minimize the error and that will give me the best field line, so there will be possibilities over here. So, I can place a line over here I can place a line over here like this. So, I can place innumerable lines like that and whichever gives me the minimum error over here that will be the best fit line like that and MINITAB does it automatically for you ok.

So, derivatives of this error square over here will give you the mathematical relationship of  $\beta_0$  and  $\beta_1$  and if we can equate it with 0 and that is the normal equation we consider and when we do that what we get is that  $\beta_0$  estimation and  $\beta_1$  estimation. So, I have a set of x condition I have set of y conditions. So, I have 1, 2 observations like this n number of observations, so every pair of observations that I get over here.

So, specifically I can calculate what is y average of that values and x average of the data set like that. So, this value will be used and  $\beta_1$  estimation is given over here. So, this is a complex 1, but this is not difficult because I know what are the values of  $x_i$ , i varies from 1 to n n number of observations that we are considering over here.

So, all these values can be calculated and  $\beta_1$  can be calculated and when  $\beta_1$  is estimated  $\beta_0$  can also be estimated. Now MINITAB does it automatically this is based on normal equation solution of normal equations over here and this is known as least square estimation this is known as least square estimation and these are unbiased estimates basically these are unbiased estimate, this is statistician what they have suggested these are unbiased estimation.

(Refer Slide Time: 05:13)

**Quality Control and Improvement using MINITAB**

**Model Adequacy**

Fitting a regression model requires several assumptions.

1. Errors have **constant variance** (Homoscedastic),
2. **Errors are uncorrelated** random variables with mean zero (No autocorrelation),
3. Errors are **normally distributed** (White Noise) A-b

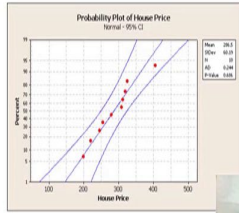
**Check Normality**

$H_0$ : The distribution is normal  
 $H_1$ : The distribution is NOT normal

**Test**

- Kolmogorov-Smirnov (K-S) Test
- Anderson-Darling (AD) Test
- Shapiro-Wilk

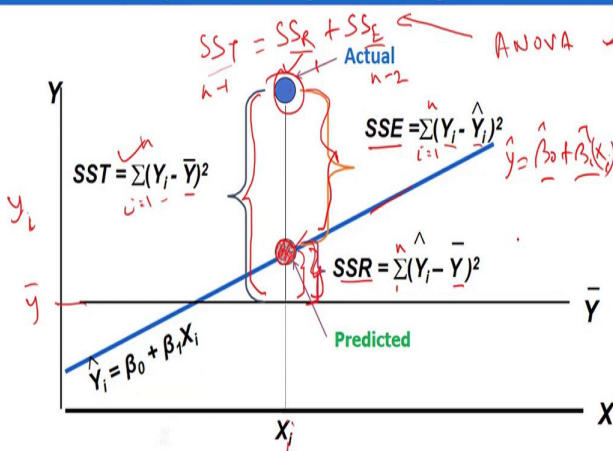
$e_i = (y_i - \hat{y}_i)$   
RESIDUAL



Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay

(Refer Slide Time: 05:14)

**Quality Control and Improvement using MINITAB**



$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$

$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$

$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

$\hat{Y}_i = \beta_0 + \beta_1 x_i$

$Y = f(X)$   
 $df = 1$

ANNOVA

Actual

Predicted

Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay

And so we can adopt that one, so in this case what we will do is that this estimation MINITAB will automatically give it for you and then there is an ANOVA analysis concept over here in MINITAB also like earlier Anova analysis over here. So, this is let us say  $y$  is on this side and  $x$  is on this side  $x_i$  values over here  $y_i$  values on this direction over here single  $x$  and single  $y$  I am just representing one point over here.

So, one actual value is located over here and based on a line equation which I can assume for a given  $\beta_0$  and  $\beta_1$ . So, this can be one of the line equation over here  $\beta_0$  plus  $\beta_1$  for a given estimation let us say and we have developed some equation and this is the line equation. So, for a given value of  $x$  that is actually  $x_i$  over here, what we can do is that we can also get a predicted value over here we can get a predicted value over here because this is the line equation.

So, whenever I put  $x_i$  in this equation I will get a predicted value which is nothing but what we are seeing over here. So, but the actual value is over here and this is the predicted value over here ok. So, this point from the overall average how much it differs.

So, this is  $y_i$  variables over here. So, this will have some average which is the  $y$  average or over average although all the  $y$  values over here. So, what is the total deviation over here? This is the total deviation that we are seeing over here.

So, this is for one observation I am saying this is the total deviation from the mean values over here and out of this how much is explained by regression equation, this is known as this is the part that is explained by regression equation. But this part of the variability is unexplained by the regression equation over here. So, the total variability from the mean overall mean over here is known as SST this is known as SST which is represented over here.

And the part of explained variability that we are seeing over here is known as SS regression, basically SST is equals to SS regression over here and the remaining which is unexplained over here that variability which we are seeing is SSE over here. Now this is for one variable one observation over here there can be  $n$  number of observations. So, this summation equals to 1 to  $n$  over here summation equals to 1 to  $n$  over here.

So, like this summation 1 to  $n$  like that so SST plus SSR plus SSE over here. So, same concept of regression again SS sum of square regression sum of square error over here

and because we are predicting  $y$  is a function of  $x$  over here,  $x$  is a single variable and the degree of freedom for a single predictor will be equals to 1 basically, so that is considered is Anova analysis.

And if I have  $n$  number of observation SST degree of freedom will be  $n$  minus 1 and the regression degree of freedom will be 1 and error degree of freedom will be  $n$  minus 2 like that. So, and that is the interpretation  $n$  minus 1 minus 1 will be  $n$  minus 2 basically ok. So, the same concept is used over here only thing is that  $X_i$  is continuous variable and it can take any values not that specific values of predefined level 1 level 2 like that that is not the case here  $X$  is continuous  $Y$  is continuous also.

So, every values I can calculate what is the SS total over here, that is the deviation from the overall average that is taken over here considered this one and this is actual values over here and this is the predicted one that will give me SS error. And the regression will be predicted minus average from the average how much is explained basically what we are getting over here.

So, this gives an idea that analysis of variance can also be adopted over here and this is used for regression analysis for adequacy model adequacy checks like that ok. And that we will see in MINITAB the interpretation of that and here also model adequacy checks are required. So, in this case constancy of variance that we have homoscedastic has to be checked. Errors are uncorrelated that means, Durbin-Watson statistics is used and normal distribution assumptions we have Anderson darling test over here to prove that one ok.

Anderson darling test means the distribution is normal, so Anderson darling test we can adapt over here. We can store the residual and residual over here is nothing but so error over here.

So, error is known as residual and that is nothing but actual values minus predicted values like that for a given observation  $x_i$  condition like that ok. So, for a given  $x_i$  what is the prediction over here? So, this is the actual  $x_i$  values over here, so this gives you the error which is known as residual which is known as a residual ok.

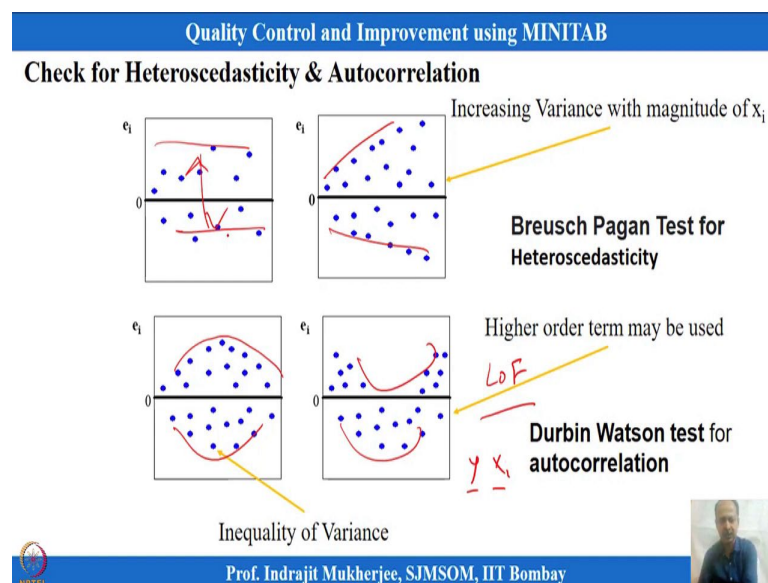
And MINITAB in MINITAB you can save the residuals and do all kind of analysis. So, sometimes residuals are stored sometimes standard residuals are stored. So, preferably we use standard standardized residual like that. So, that takes care of certain other

aspects like that. So, what we will do is that we will we will just analyze the residual and try to see model adequacy check.

So, this is important over here this is the assumptions has to be satisfied to use the regression model which can be, if model assumptions are satisfactory immediately we can say it can be generalized and irrespective. If you give me the within the range of  $x$  if you give me any values I can predict. Even if I do not have the observation in earlier in the historical data set any new observation within that boundary conditions of  $x$  I can predict what will be the expected value of  $y$  basically ok.

So, prediction is possible of  $y$  for a given value of  $x$  ok, only thing I have to understand that it should be within the boundaries where the regression model is developed. So, I cannot extrapolate basically regression equations. So, that consideration is always there ok.

(Refer Slide Time: 10:24)



So, heteroscedasticity we have already discussed that if this is the funnel shape in that case heteroscedastic behavior, if this is the scenario then it says that linear model may not be sufficient you may have to incorporate like second order terms or something like that.

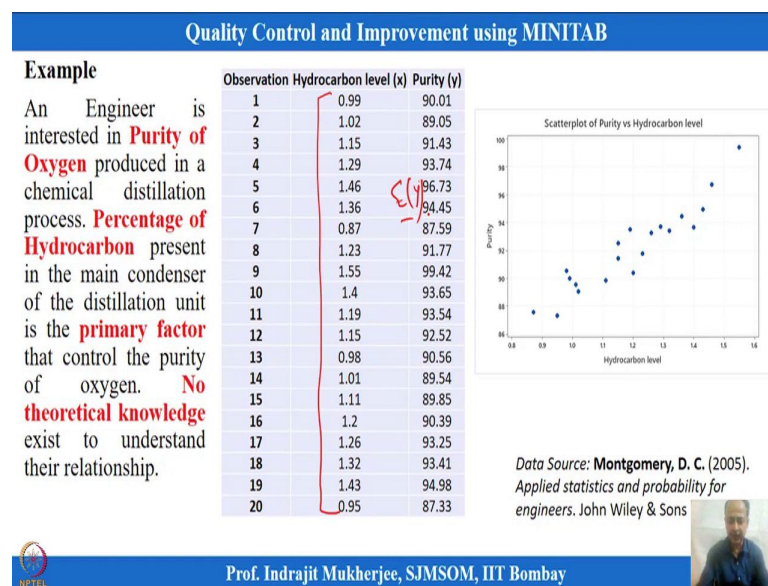
So, this is also lack of fit this will be reflected in lack of fit which is known as lack of fit integration. So, this if you have multiple observations like  $x_i$  over here and  $y_i$



observations over here. And if you have x repetition observation for a given x there are at a given condition of x I have multiple observations like that, then only I can calculate this lack of fits like that.

So, that will give me nonlinearity whether it is non-linearity is there or not this is also non constancy of variance over here and this is more or less what is expected. So, there is no as such deviations that is happening, so that is residual versus fit we can plot that one and see that one ok.

(Refer Slide Time: 11:19)



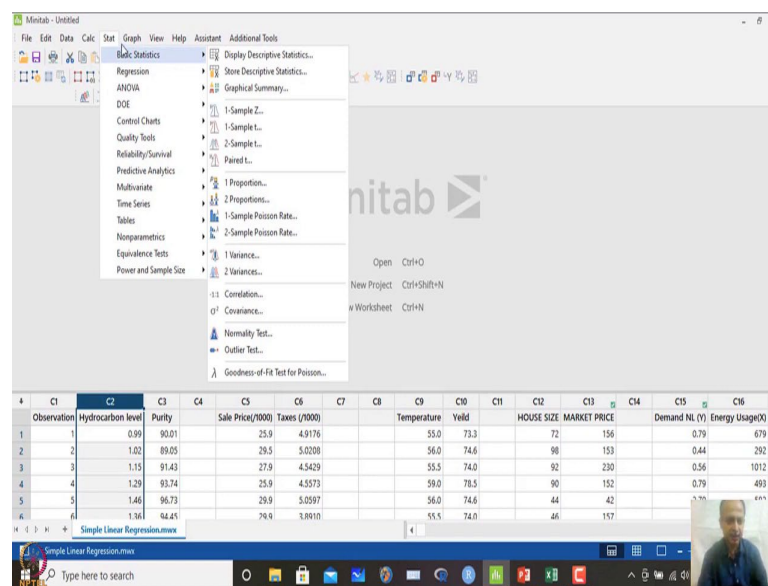
So, if that is the scenario if this is the scenario again that conversion or transformation has to be used on y basically and then regress transform y with the x variables like that ok. So, here we are taking one example to illustrate regression in MINITAB. So, for this what we will do is that this is an example an engineer is interested this is taken from one of the mode is applied statistics and probability for engineers, this is the data set that that is given.

So, purity of data purity of oxygen over here is basically y variable and percentage of hydrocarbon is considered as one of the factor this is the historical data not based on experimentation statistical experimentation. But just historical data which says and I want to; I want to check whether the hydrocarbon levels, when I change whether it is expected value of purity is changing and whether I can develop a generalized equation prediction equation and within this range that is given over here.

Whether I can predict the expected value of y for a whether it is possible or not by developing a regression equation which can be generalized like that, if all model adequacy checks are ok then in that case we can do that ok.

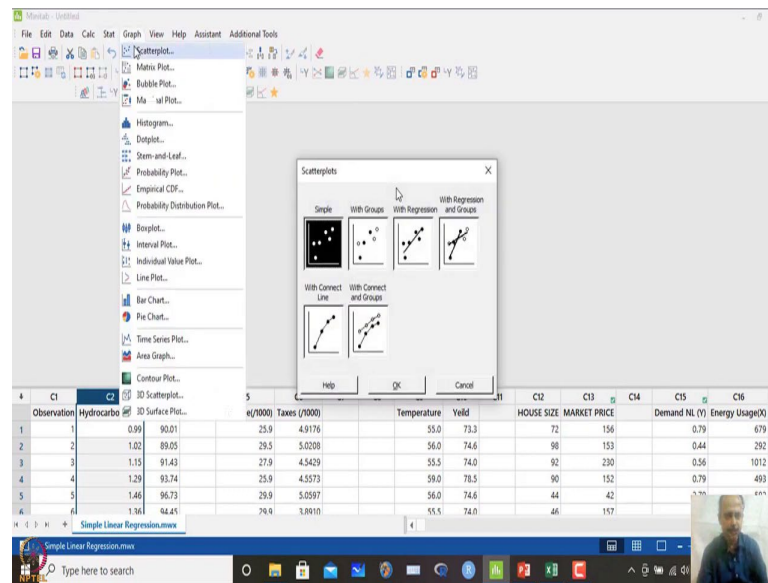
So and why we are doing this because why we are doing this because we do not have this we do not have this theoretical functions that can be used to model purity with hydrocarbon levels basically. And so that is one of the; one of the constraints that we have, that is why we are doing empirical modeling over here ok.

(Refer Slide Time: 12:57)

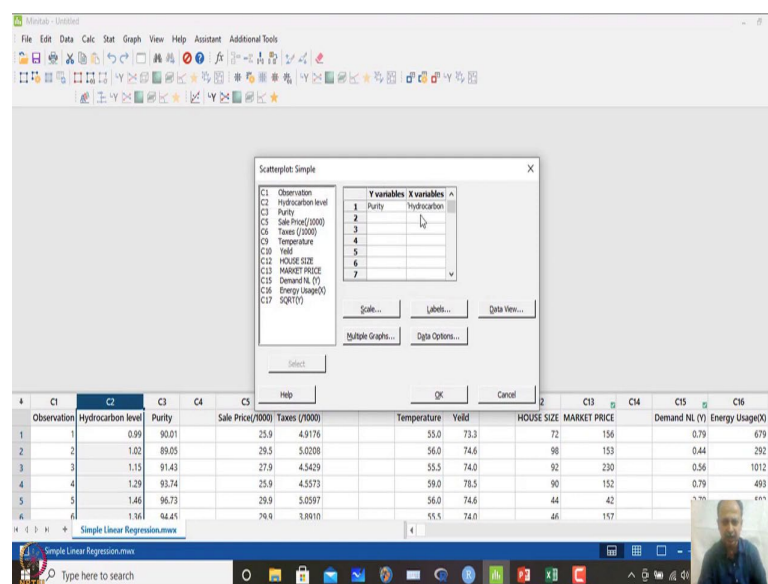


So, how do we do that that is important for us? So, I will just go to that interface of MINITAB and this is the C1, C2 observation Ct observation that is given over here. I have just taken the data 1 is purity data 1 is hydrocarbon data over here and in this case what I have to do is that I have to go to stat and regression analysis. So, let us just try to see whether scatter plot what does caterpillar shows over here.

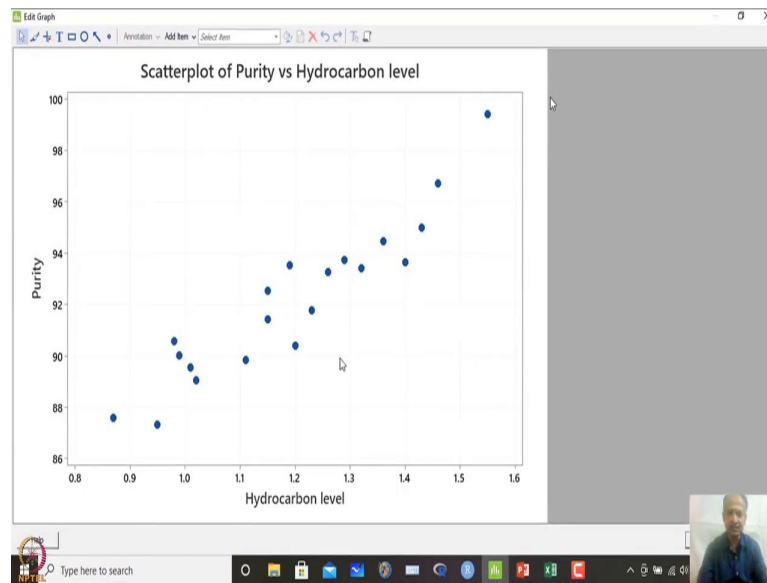
(Refer Slide Time: 13:19)



(Refer Slide Time: 13:23)



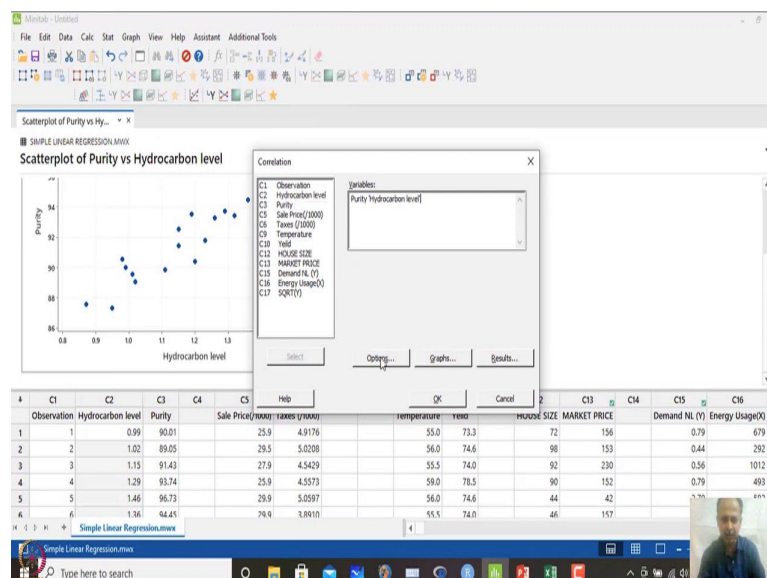
(Refer Slide Time: 13:30)



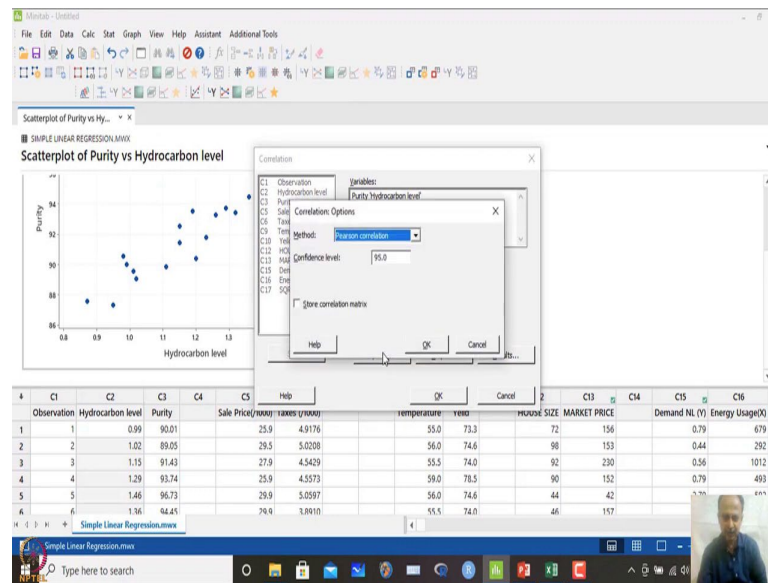
So, whether we can see an see that linear relationship or not. So, you can go to graph and scatter plot. So, in this case what we can do is simple scatter plot over here and y variable is purity over here and x variable is hydrocarbon.

And if you click ok over here what will happen is that it will give you some this is the graph that we that we can see over here and it shows that there is a linear relationship that exists over here and also we can check the correlation. So, this is positive correlation what we can see is that hydrocarbon increases purity also increases.

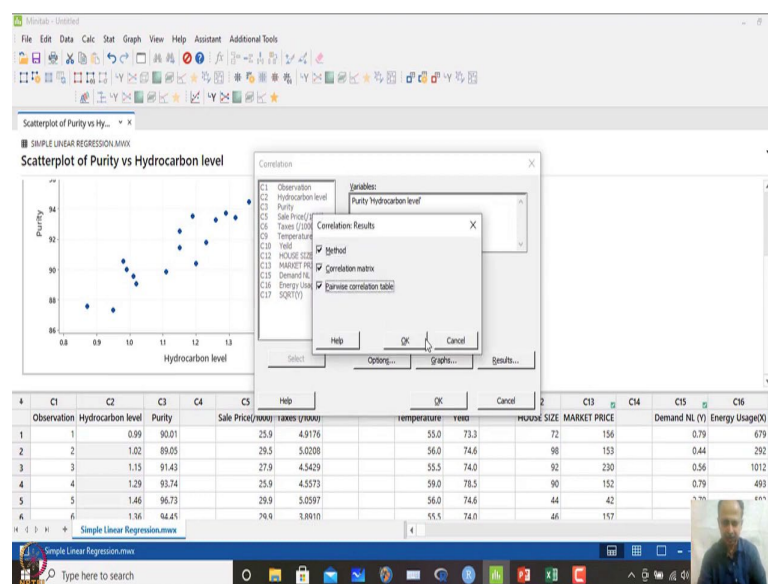
(Refer Slide Time: 13:50)



(Refer Slide Time: 13:56)

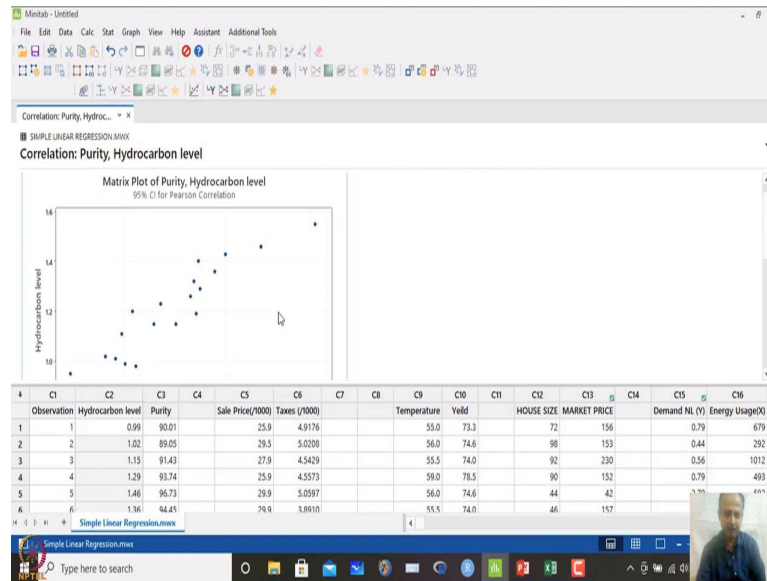


(Refer Slide Time: 14:02)

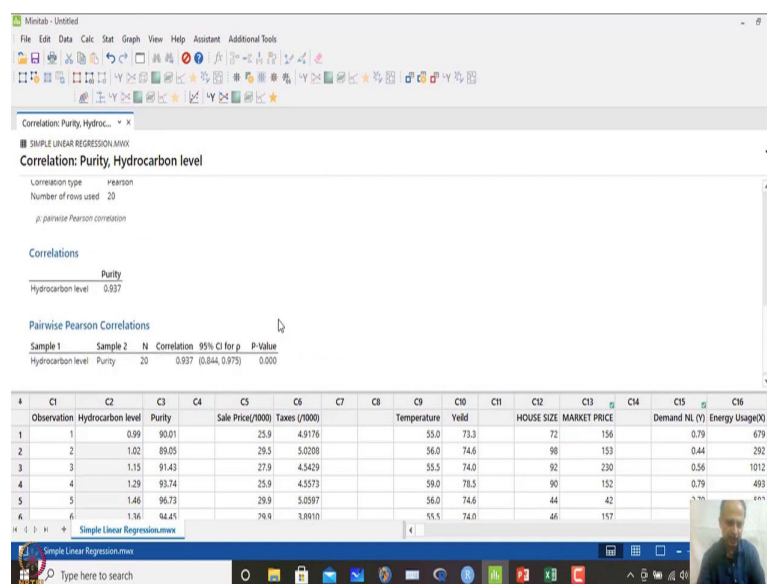


So, we can just check the correlation over here. So, basic statistics we can just check correlations and correlations I want to see purity with hydrocarbon levels over here and in options we can just do this and results we can just interpret this one and I click ok.

(Refer Slide Time: 14:04)



(Refer Slide Time: 14:07)



And what we see is that correlation p-values that we are seeing over here is near to 0 and less than 0.05. So, that indicates that this and purity is basically highly correlated and correlation value what you can see is that 0.937 and that is more than 0.7 and that is significant over here 0.93 is very good ok.

So, a p-value indicates that there is a statistical significance over here and the p-value is quite p-value shows significance over here; that means, they are highly correlated and linearly correlated basically. So, Pearson correlation is used over here.



(Refer Slide Time: 14:42)

Stat > Regression > Fitted Line Plot...

Correlation: Purity, Hydrocarbon level

Pairwise Pearson Correlations

Sample 1	Sample 2	N	Correlation	95% CI for $\rho$	P-Value
Hydrocarbon level	Purity	20	0.937	(0.844, 0.973)	0.000

Simple Linear Regression

Observation	Hydrocarbon level	Purity	Sale Price(1000)	Taxes (1000)	Temperature	Yield	HOUSE SIZE	MARKET PRICE	Demand NL (Y)	Energy Usage(X)	
1	1	0.99	90.01	25.9	4.9176	55.0	73.3	72	156	0.79	679
2	2	1.02	89.05	29.5	5.0208	56.0	74.6	98	153	0.44	292
3	3	1.15	91.43	27.9	4.5429	55.5	74.0	92	230	0.56	1012
4	4	1.29	93.74	25.9	4.5573	59.0	78.5	90	152	0.79	493
5	5	1.46	96.73	29.9	5.0597	56.0	74.6	44	42	0.79	493
6	6	1.36	94.45	29.4	3.8610	55.5	74.0	46	157	0.79	493

(Refer Slide Time: 14:48)

Correlation: Purity, Hydrocarbon level

Pairwise Pearson Correlations

Sample 1	Sample 2	N	Correlation	95% CI for $\rho$	P-Value
Hydrocarbon level	Purity	20	0.937	(0.844, 0.973)	0.000

Regression

Responses: Purity

Continuous predictors: Hydrocarbon level

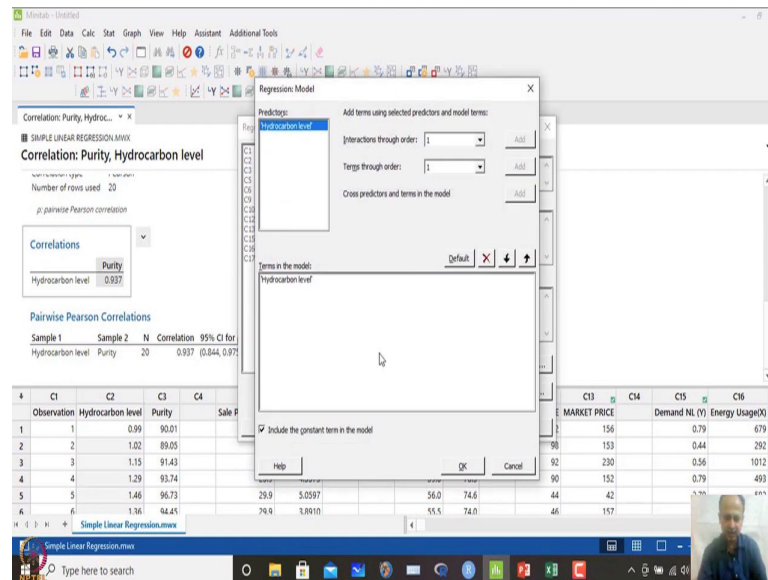
Categorical predictors:

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

So, if this is so then I can develop the regression equation. So, let us develop the regression equation. So, what I do is that I go to stat I go to regression and I go to fit regression models over here, then it will ask what is the response? I will say purity is the response, what is the continuous predictor I have hydrocarbon there is no categorical predictor over here that is also possible to incorporate.

So, we will not go into that complexity now ok. So, if purity is the response and hydrocarbon is the only x variable that is continuous and then all these options there are many more options over here what you can see.

(Refer Slide Time: 15:09)



So, only thing what we will consider include the constant term in the model. So, this should be clicked over here, because statistician says that when I include the intercept when I include the constant term which is the intercept over here, generally the model performance is quite good and this is seen by many research like that.

So, we will not omit this  $\beta_0$  estimation over here we will keep that one. So, this is the only thing that you have to; you have to remember, these options coding stepwise that that is not required at this stage validation graphs results and only thing what you can do is that you can store the residuals.



(Refer Slide Time: 15:43)

The screenshot shows the Minitab software interface. A 'Regression' dialog box is open, with 'Standardized residuals' selected under the 'Residuals' section. The background window displays the 'Correlation: Purity, Hydrocarbon level' output, showing a Pearson correlation of 0.937. Below this, a table of data is visible, including columns for Observation, Hydrocarbon level, Purity, and Sale Price.

Observation	Hydrocarbon level	Purity	Sale Price
1	0.99	90.01	29.5
2	1.02	89.05	27.9
3	1.15	91.43	27.9
4	1.29	93.74	25.9
5	1.46	96.73	29.9
6	1.36	94.45	29.9

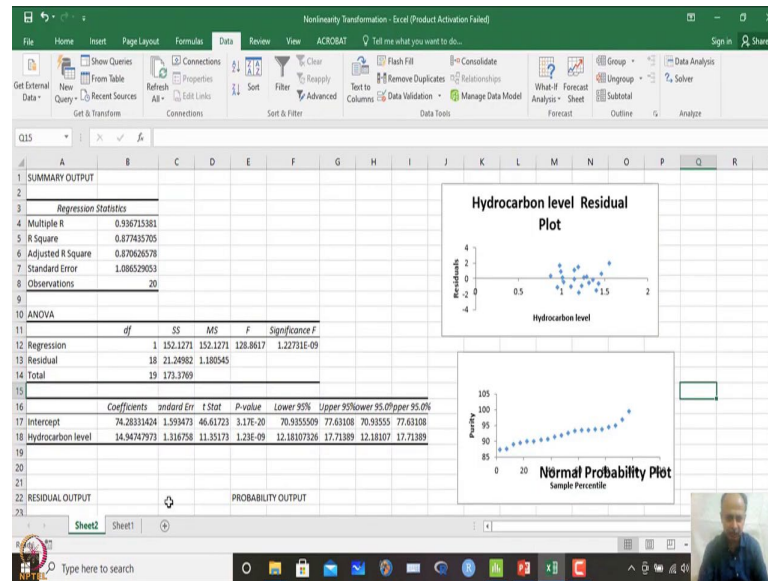
Now, I can store the residuals of standardized residual let us store, standardized residual which is generally recommended and this is nothing but residuals divided by standard deviation of residuals. So, that is known as standardized residual and we want to save that one and when I save this one and I click ok.

(Refer Slide Time: 16:09)

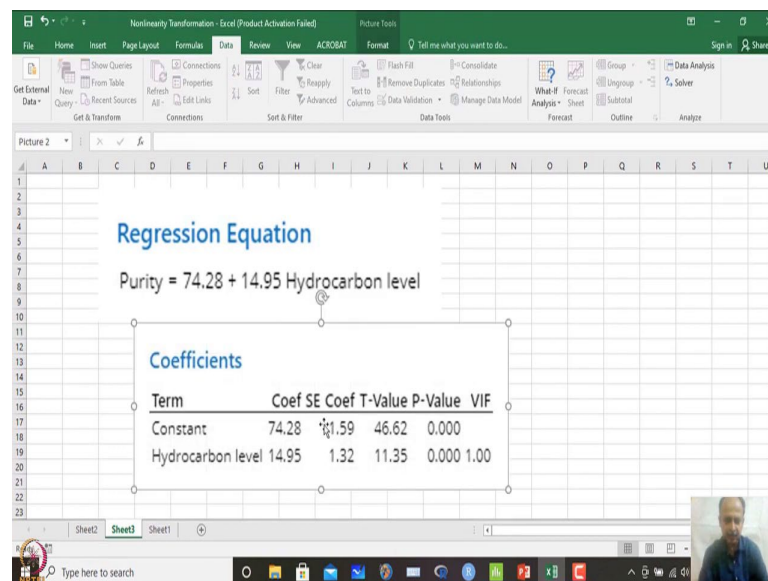
The screenshot shows the Minitab software interface. The 'Regression Analysis: Purity versus Hydrocarbon level' output is displayed. The 'Regression Equation' is shown as  $Purity = 74.28 + 14.95 \text{ Hydrocarbon level}$ . Below this, the 'Coefficients' table is visible, showing the coefficients for the regression equation. A context menu is open over the 'Coefficients' table, with 'Copy Picture' selected.

Term	Coef	SE Coef	T-Value
Constant	74.28	1.59	46.62
Hydrocarbon level	14.95	1.32	11.35

(Refer Slide Time: 16:16)



(Refer Slide Time: 16:20)



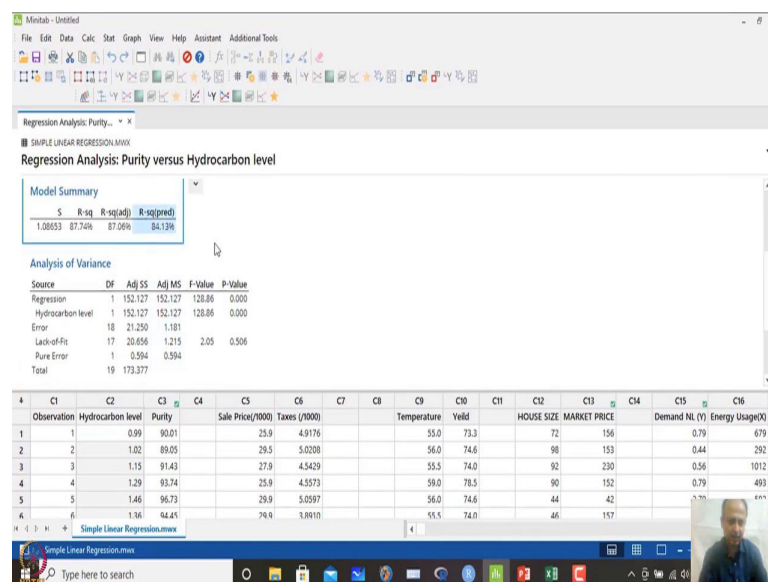
What will happen is that I will get the regression equation. When I get the regression equation, so this is the equation that you are seeing. So, I can copy this one copy as picture and I can just paste this one. Let us say I want to see the results. So over here what I have done is that this is the first result, the equation that MINITAB has generated is purity is equals to this is the intercept  $\beta_0 = 74.28$  this is estimated based on the formulation that we have shown.

And 14.95 is the  $\beta_1$  that is the slope and 1 unit increase of hydrocarbon level, how much will be the average increase in the purity that is given by 14.95 like that. So,  $\beta_0$  estimation  $\beta_1$  estimation over here and then if you go to the second results these are the this is the second results that you will get and which we can paste over here and see this one.

And here the coefficient is given, so when I see constant that is the  $\beta_0$  estimation coefficient is 74.28 which is reflected over here and hydrocarbon is  $\beta_1$  intercept is 14.95 that is given over here ok. Standard error of this is also given and the corresponding T values and P-values are given P less than 0.05 will indicate that  $\beta_1$  is significant over here and this is statistically significant.

So, it is basically saying that there is a slope and we can consider this hydrocarbon level as one of the variables which is explaining the variability of purity over here ok. Constant is also  $\beta_0$  is significant statistically, so we should retain this one  $\beta_0$  ok. So, that is also signifies signified over here.

(Refer Slide Time: 17:36)



Regression Analysis: Purity versus Hydrocarbon level

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.08053	87.34%	87.06%	84.13%

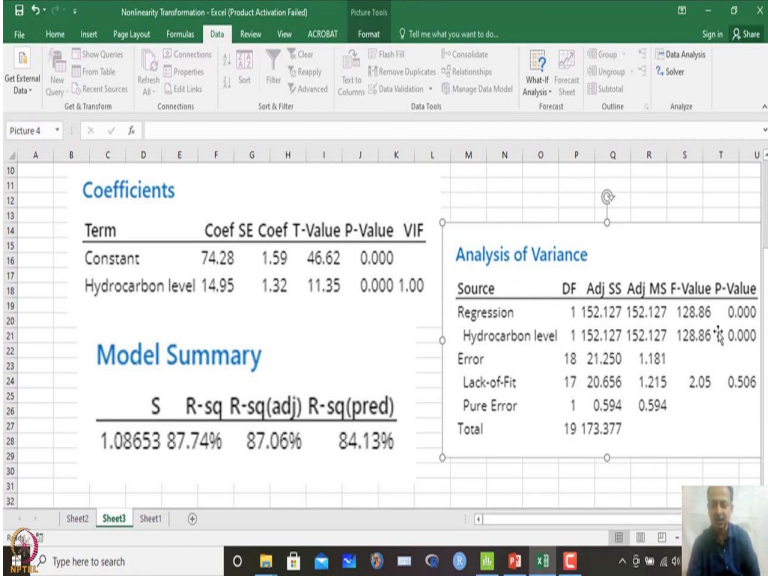
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	152.127	152.127	128.86	0.000
Hydrocarbon level	1	152.127	152.127	128.86	0.000
Error	18	21.250	1.181		
Lack-of-Fit	17	20.656	1.215	2.05	0.506
Pure Error	1	0.594	0.594		
Total	19	173.377			

Observation	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
	Hydrocarbon level	Purity		Sale Price/(1000)	Taxes/(1000)			Temperature	Yield		HOUSE SIZE	MARKET PRICE			Demand NL (Y)	Energy Usage(X)
1	1	0.99	90.01		25.9	4.9176			55.0	73.3		72	156		0.79	679
2	2	1.02	89.05		28.5	5.0208			56.0	74.6		98	153		0.44	292
3	3	1.15	91.43		27.9	4.5429			55.5	74.0		92	230		0.56	1012
4	4	1.29	93.74		25.9	4.5573			59.0	78.5		90	152		0.79	493
5	5	1.46	96.73		28.9	5.0597			56.0	74.6		44	42			
6	6	1.36	94.45		28.8	3.8010			55.5	74.0		46	157			

And then model summary you will find over here model summary. So, I can just copy this one also copy as picture and we can press this one ok.

(Refer Slide Time: 17:48)



**Coefficients**

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	74.28	1.59	46.62	0.000	
Hydrocarbon level	14.95	1.32	11.35	0.000	1.00

**Model Summary**

S	R-sq	R-sq(adj)	R-sq(pred)
1.08653	87.74%	87.06%	84.13%

**Analysis of Variance**

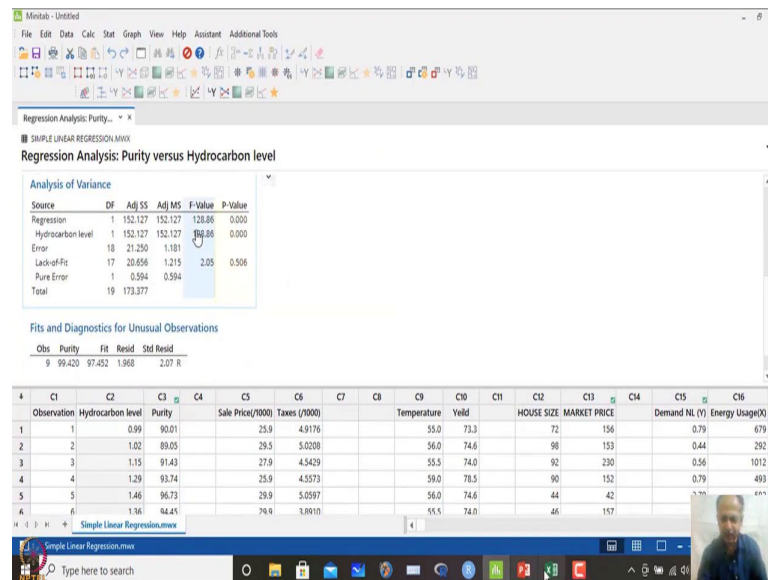
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	152.127	152.127	128.86	0.000
Hydrocarbon level	1	152.127	152.127	128.86	0.000
Error	18	21.250	1.181		
Lack-of-Fit	17	20.656	1.215	2.05	0.506
Pure Error	1	0.594	0.594		
Total	19	173.377			

So, we can place this one over here, so in this case model adequacy. So, this is one of the; one of the measures that like R values that we have told correlation coefficient is nothing but the correlation coefficient that you see over here R square which is known as coefficient of determination, which is same in case there is one variable over here. So, it will come out to be same 87.74. So, it is converted into percentage.

So, I can convert into between minus 1 and plus 1, so 0.877 you can think of this is more than 0.7 ok and this is calculated by another formulas over here that is known as SS regression. How much of the variability by SS total basically which Anova analysis will tell you and the Anova tables will summarize that one. So, SS regression but SS total will tell me how much of the variability of y is explained by basically this hydrocarbon level variable over here.

So, it is around 87 percent like that so you can think of ok. So, that is quite good one of the variable is explaining so much of variability of the y that you observe. So, when I change the x it is influencing the expected value of y basically that is the interpretation that we can make out of this ok.

(Refer Slide Time: 18:55)



So, then what are the other results that we are getting over here this is the analysis of variance that you see over here. So, this I can copy again and I can paste it over here. So, I can paste it here and just try to see what is the interpretation of this ok. So, over here what you see is that regression this is the regression equation that is developed, it is showing that P-value is less than 0.05.

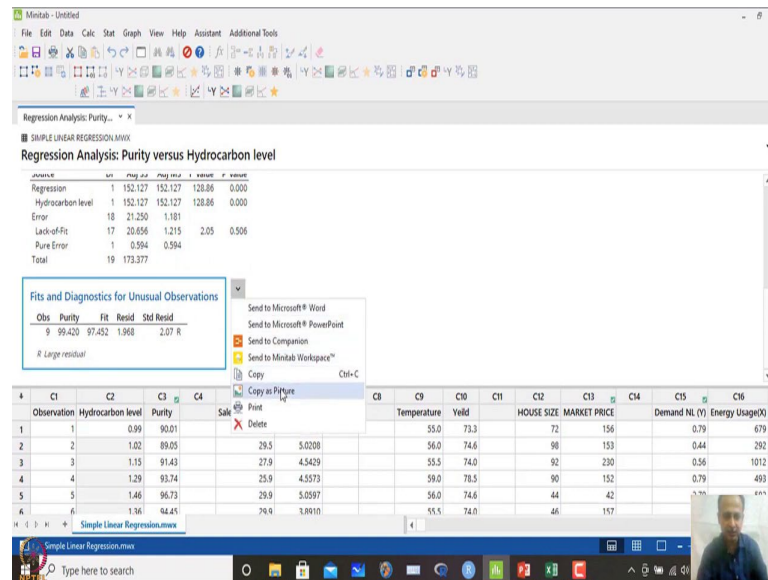
That means, this regression is quite this equation is quite significant and in that case we can adopt this equation and hydrocarbon level is that is basically the variable x that we are considering P-value is significant over here. And we will also find a lack of fit testing over here, that means whether there is any nonlinearity in the model, that is we have to adopt and go to higher order equations like that, that will be given by lack of fit test and a formula is given in any books like that.

So, if you have multiple observation at a given level of x then lack of fit can be calculated and lack of fit over here is calculated as 0.5 which is not more than less than 0.05; that means, there is no lack of fit as such. So, linear model is quite sufficient to explain the variability and that is adequate over here. So, I do not need to go to higher order terms over here.

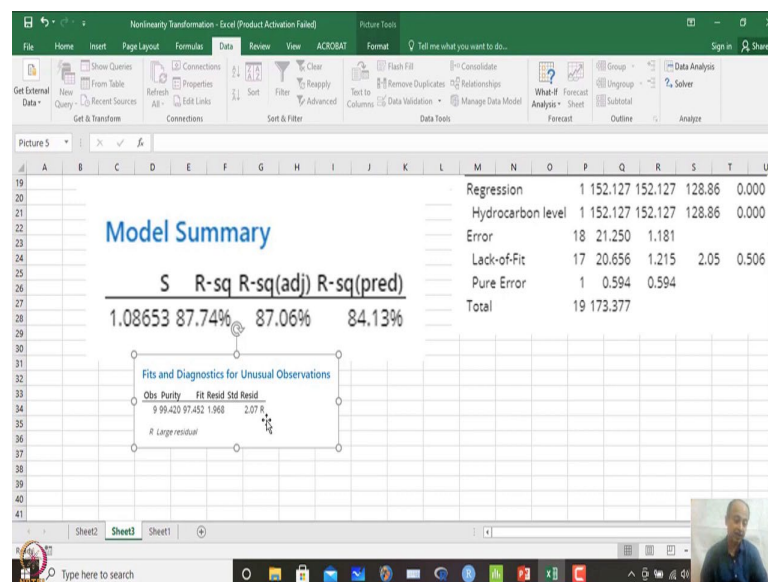
So, this is lack of fit what we can get ok. So, this regression equation one way is seeing these coefficients over here  $\beta_1$  is significant  $\beta_0$  is significant like that and overall if you want to see that whether the regression equation is making sense.

So, this value of regression this P-value we have to see and generally it will agree. So, both of the coefficient is significant then only regression will be significant like that, so that is the interpretation that we can make ok.

(Refer Slide Time: 20:32)



(Refer Slide Time: 20:42)



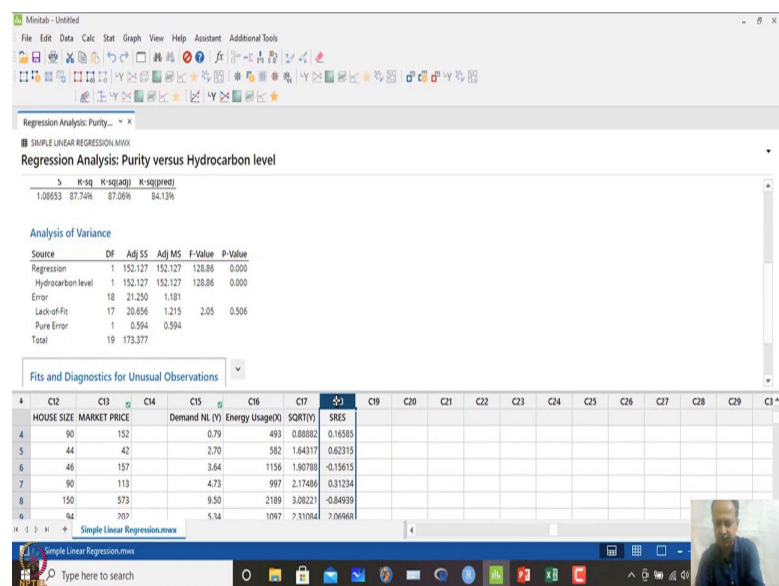
Then some unusual observations over here when standard residual is more than. So if you can see this one copy as picture; so we can press this one. So, when we have certain observation which is beyond two that is standardized residual which indicates that this is the unusual observations like that.



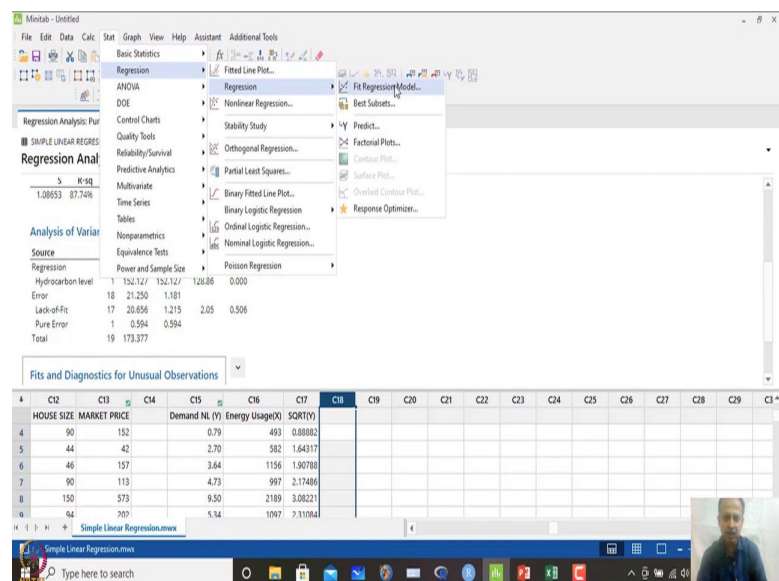
So, then we have to see whether to include that one or exclude that one. So, whether it is outlier like that whether we need to eliminate that one, regression also this type of information is also useful when we do regression analysis. So, that is why standardized residual is used to identify any outlier observations something we can do that ok.

So, and we have to be careful in dealing with outliers, so there are many ways of dealing with outliers ok. So, what we can understand based on this simple example over here is that every condition is satisfactory.

(Refer Slide Time: 21:22)



(Refer Slide Time: 21:31)



(Refer Slide Time: 21:32)

Regression Analysis: Purity versus Hydrocarbon level

Regression equation:

$$\text{Purity} = 1.08553 + 0.00000 \text{ Hydrocarbon level}$$

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	152.127	152.127	128.86	0.000
Hydrocarbon level	1	152.127	152.127	128.86	0.000
Error	18	21.250	1.181		
Lack-of-Fit	17	20.656	1.215	2.05	0.051
Pure Error	1	0.594	0.594		
Total	19	173.377			

Regression dialog box:

- Response: Purity
- Continuous predictors: Hydrocarbon level

(Refer Slide Time: 21:34)

Regression Analysis: Purity versus Hydrocarbon level

Regression equation:

$$\text{Purity} = 1.08553 + 0.00000 \text{ Hydrocarbon level}$$

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	152.127	152.127	128.86	0.000
Hydrocarbon level	1	152.127	152.127	128.86	0.000
Error	18	21.250	1.181		
Lack-of-Fit	17	20.656	1.215	2.05	0.051
Pure Error	1	0.594	0.594		
Total	19	173.377			

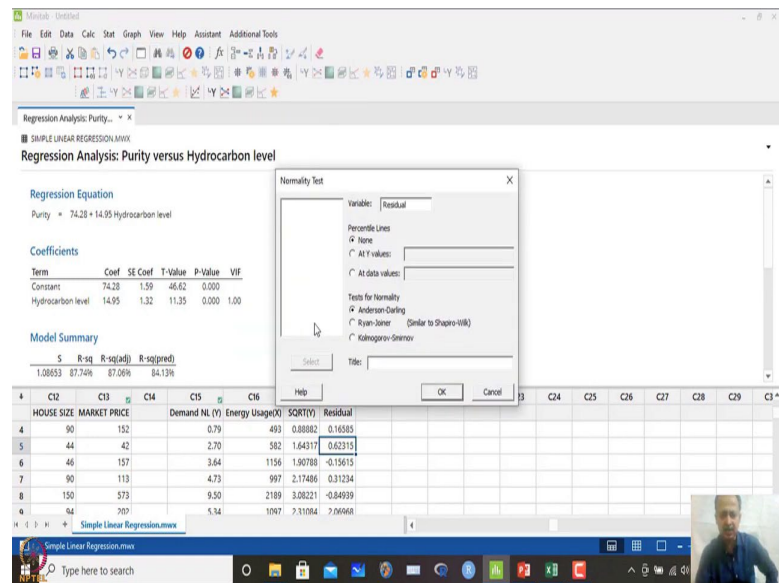
Regression Storage dialog box:

- ☒ Residuals
- ☒ Standardized residuals
- ☒ Deleted residuals
- ☒ Leverages
- ☐ Cook's distance
- ☐ DFFITS

Now, one more thing what we have to do is that we have to save the residuals. So, let me just delete this one, I do not know whether this is we have saved or not. So, we I go to residual regression over here fit regression like that this is the one storage whether we have standardized residual were saved over here.

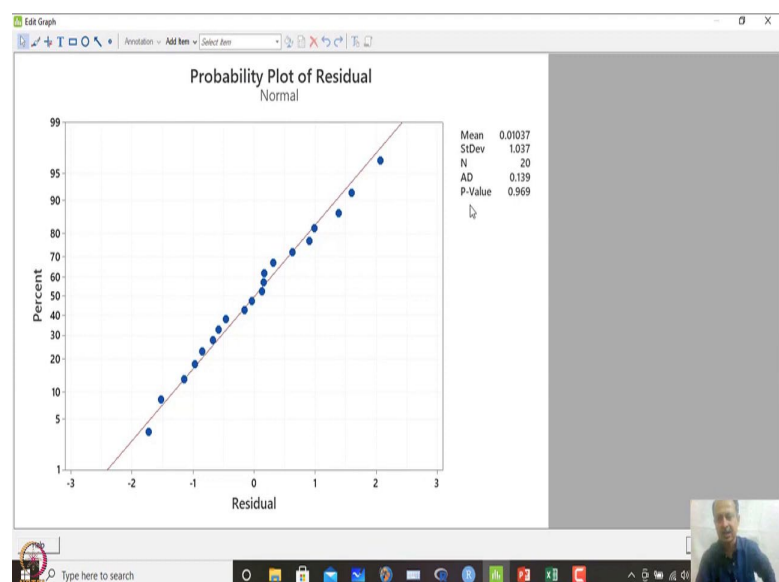


(Refer Slide Time: 21:42)



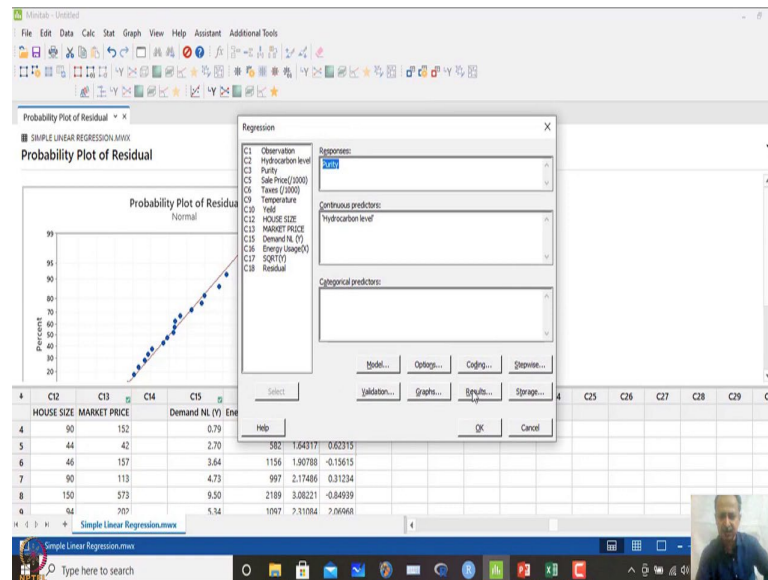
So, this is ok and if I save this one the last one is standardized residually this can be considered as residual over here and there are 3 checks that has to be under we have to undergo 3 test over here which is the one of the test is normality assumptions like that, so one test what we can do is that we can see the basic statistics we can go to normality test over here and we can see the residual whether it is normal or not.

(Refer Slide Time: 22:02)

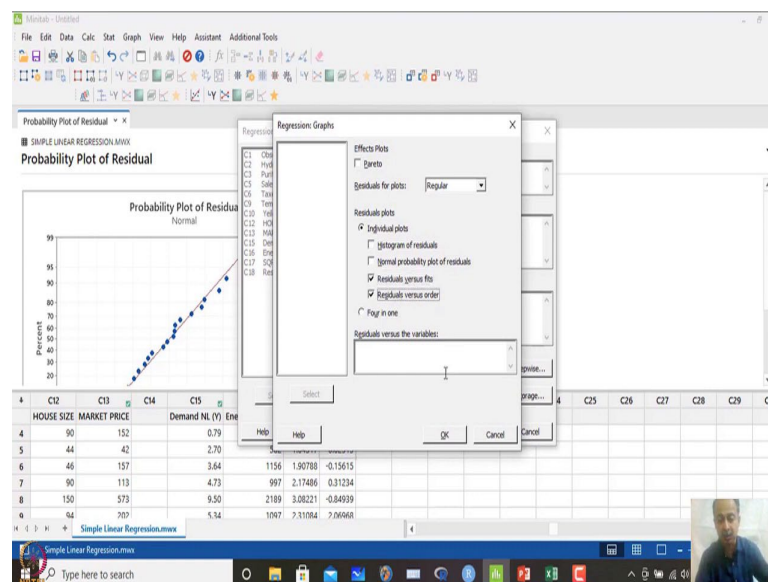


So, in this case what we observe is that P-value is more than 0.05. So, there is no problem in the assumptions of normality over here there is another graph which can be seen over here. So, when I am; when I am doing this regression.

(Refer Slide Time: 22:15)



(Refer Slide Time: 22:17)



There are other possible graphs that we can draw. So, one is residual versus fit, so this will indicate heteroscedasticity is there or whether there is any autocorrelation that exists this is the second one with respect to order.

(Refer Slide Time: 22:28)

**Regression Analysis: Purity versus Hydrocarbon level**

**Regression Equation**  
Purity = 74.28 + 14.95 Hydrocarbon level

**Coefficients**

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	74.28	1.59	46.62	0.000	
Hydrocarbon level	14.95	1.32	11.35	0.000	1.00

**Model Summary**

S	R-sq	R-sq(Adj)	R-sq(Pred)
1.08553	87.74%	87.06%	84.13%

The bottom part of the window shows a worksheet with columns C2 through C30. The data includes HOUSE SIZE, MARKET PRICE, Demand NL (Y), Energy Usage(X), SQR(Y), Residual, and SRES.

(Refer Slide Time: 22:29)

**Regression Analysis: Purity versus Hydrocarbon level**

**Analysis of Variance**

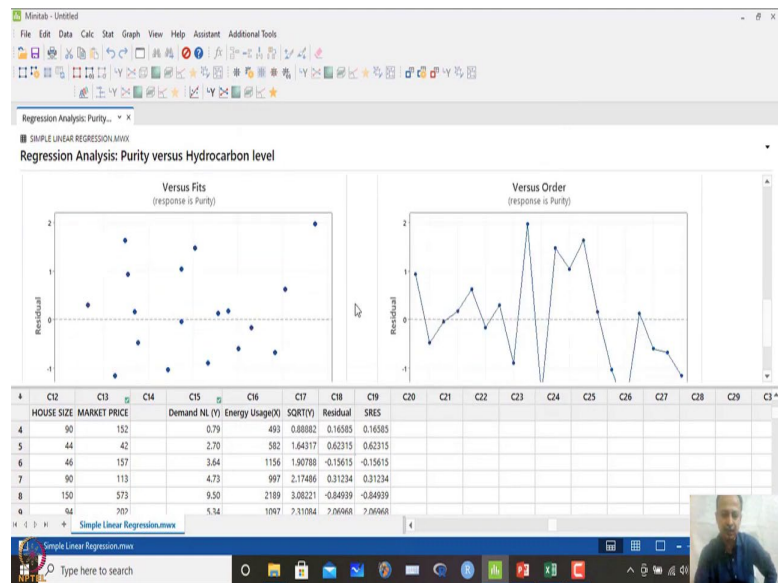
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	152.127	152.127	128.86	0.000
Hydrocarbon level	1	152.127	152.127	128.86	0.000
Error	18	21.250	1.181		
Lack-of-Fit	17	20.856	1.215	2.05	0.506
Pure Error	1	0.594	0.594		
Total	19	173.377			

**Fits and Diagnostics for Unusual Observations**

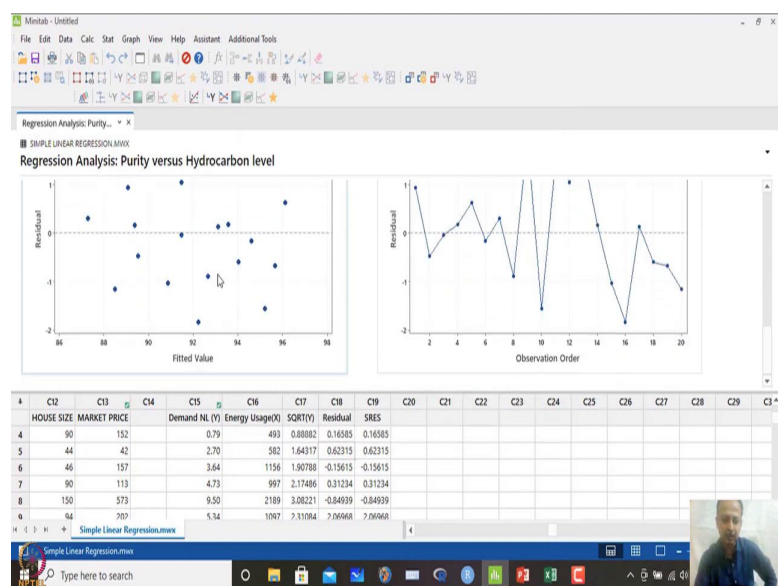
Obs	Purity	Fit	Resid	Std Resid
9	99.420	97.452	1.968	2.07 R

The bottom part of the window shows the same worksheet as the previous slide, with columns C2 through C30.

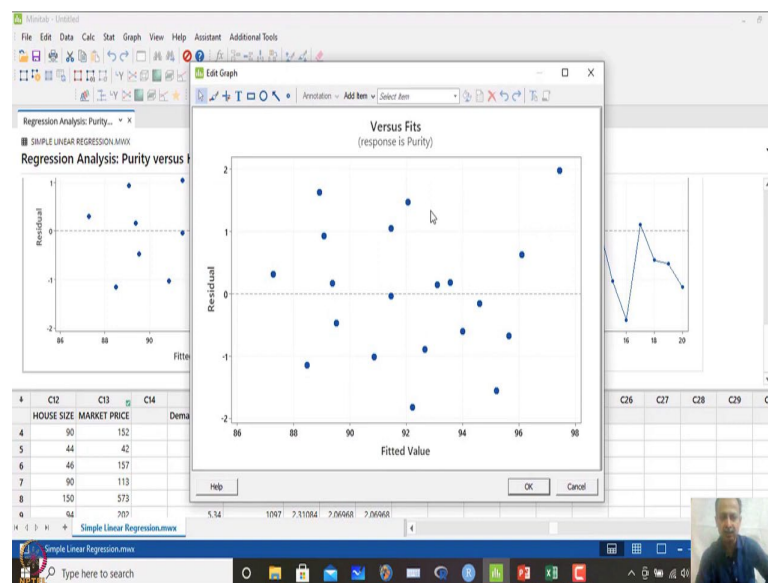
(Refer Slide Time: 22:29)



(Refer Slide Time: 22:30)

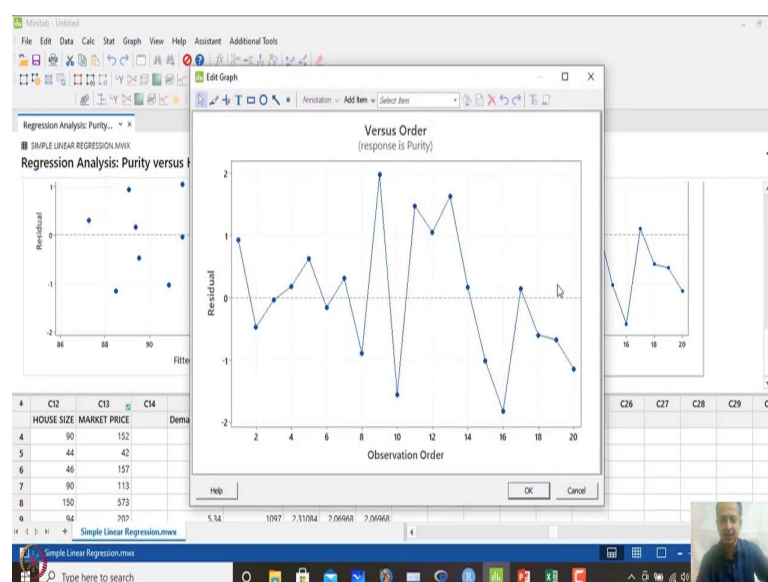


(Refer Slide Time: 22:31)



So, when you draw these two graphs also what you will observe is that this is one of the graph which will more or less you see residuals with fitted value is more or less random over here on the 0 line. So, there is no heteroscedasticity as such, but we can prove that by Breusch pagan test like that. So, we can take the residuals and we can test that one and that is available in R what I have told earlier also. So, this can be done.

(Refer Slide Time: 22:49)

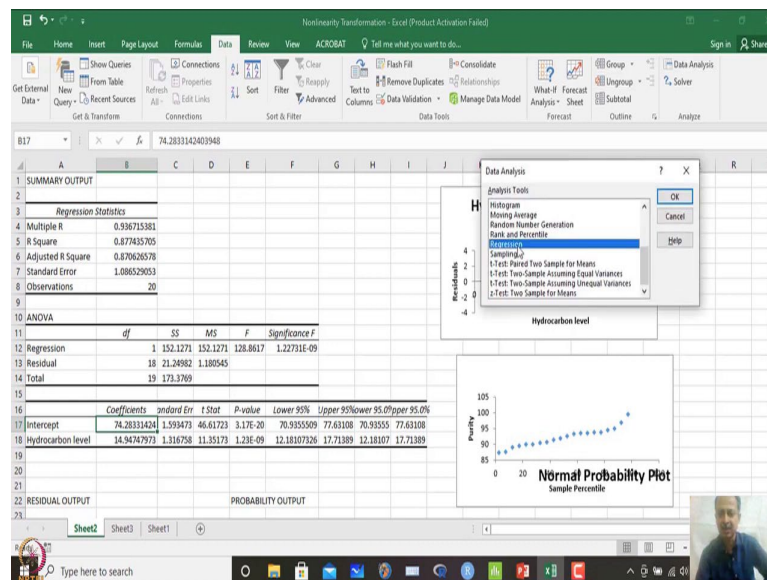


And if there is any pattern that we observe in this data set like that, but it seems to be random. So, in this case autocorrelation may also be come out to be negligible over here.

So, this is also not. So, all the assumptions we can check, but what we are seeing is that at least preliminary assumptions are satisfactory and this if you go to this book you will find that all the assumptions are true in this case and so we can generalize the equation.

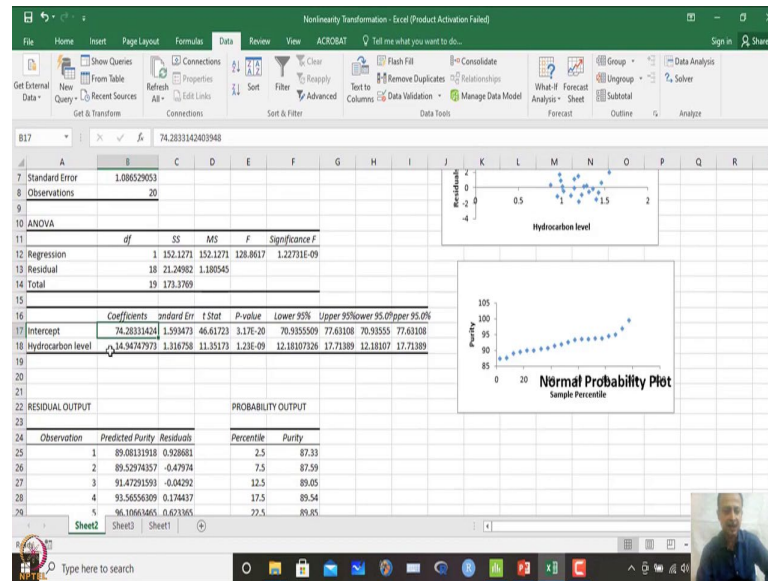
That means, this equation that we have written over here can be generalized and the values that you are getting 74.28 and 14.95 of the  $\beta_0$  and  $\beta_1$  estimation that is also possible in excel. So, if you go to excel over here and you run the regression equation and which I have done earlier over here save this one and what happens is that the values that you see constant values of intercept that you see over here by using excel simple excel.

(Refer Slide Time: 23:37)





(Refer Slide Time: 23:45)



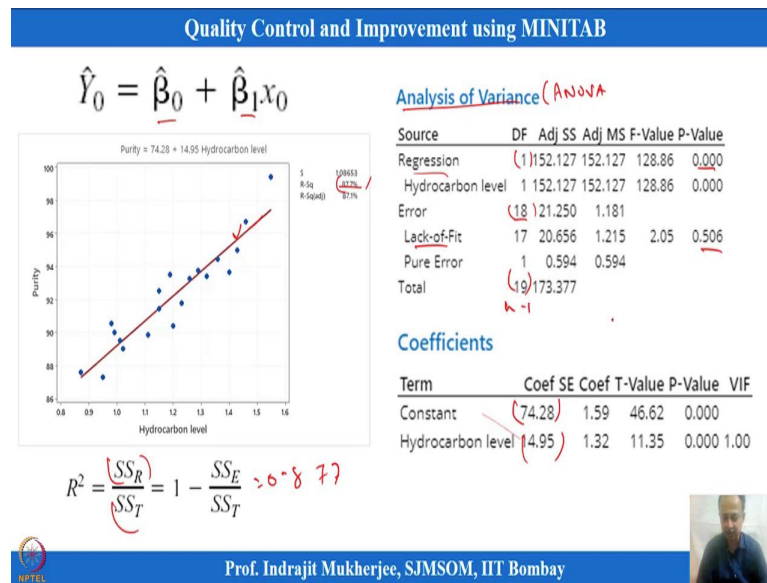
So, what I do data and I have data analysis tools and in this case you have regression analysis tools and based on that you can do the regression analysis and you can get the coefficients and corresponding P-values are also indicated over here regression is significant or not.

So, this is more than three place of decimal over here what you can see 10 to the power minus 9 and it is showing it will give you the values that is a this is not 0 what we can see over here and all the MINITAB reports this has 0. So, this MINITAB does not report more than three plus or decimal over here ok. But excel reports beyond three place of decimal over here so this is true and the values are exact what you see intercepts is 74.28 and MINITAB also has given 74.28.

So, if you go to the original value 74.28 14.95 is the  $\beta_1$  estimation that you see and excel also shows the  $\beta_1$  estimation 14.95 approximately that is also over here what you see 14.95. So, regression is also possible you can verify that one if you want to see what is the exact P-values like that and MINITAB is not giving you that because it does not go beyond 3 place of decimal.

So, in that case what you can do is that you can do it in excel and see that one or you can transfer this to R and do the analysis and you will get the exact P-values what is what comes out from the analysis like that ok. So generalized equation that we can use is 74.28 14.95 and this is hydrocarbon level that we can take ok.

(Refer Slide Time: 25:11)



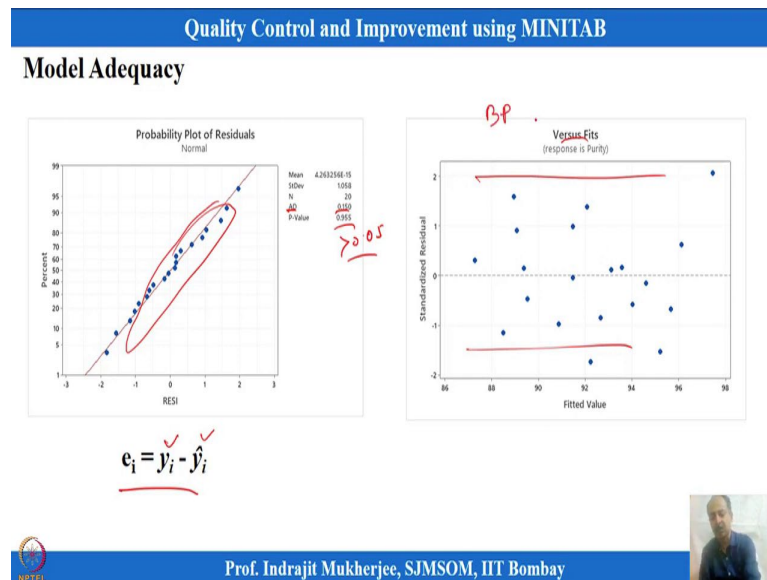
So, this is one of the example that we have seen and the data set is this is the data set and these are the graphs and equations over here. So, this is the model fit that you can see, this is the line equation that is denoted over here R square is around 0.87 7 like that. This  $\beta_0$  and  $\beta_1$  is estimated over here.

So, this is  $\beta_0$  and this is  $\beta_1$  estimation over here and this regression is found to be significant and lack of fit there is no lack of fit. So, in this case linear equation is sufficient and R square value is SS regression by SST which is approximately equals to 0.877.

What you are seeing over here 0.877 this is the value over here ok. So, this is the Anova analysis that you see analysis of variance which is the Anova analysis that you see and regression degree of freedom is 1, because 1 variable is there ok and error degree of freedom is basically n minus 2. So 20 observations are there so 20 minus 2 is 18 and n minus 1 observation is over here, so this is 19 over here.



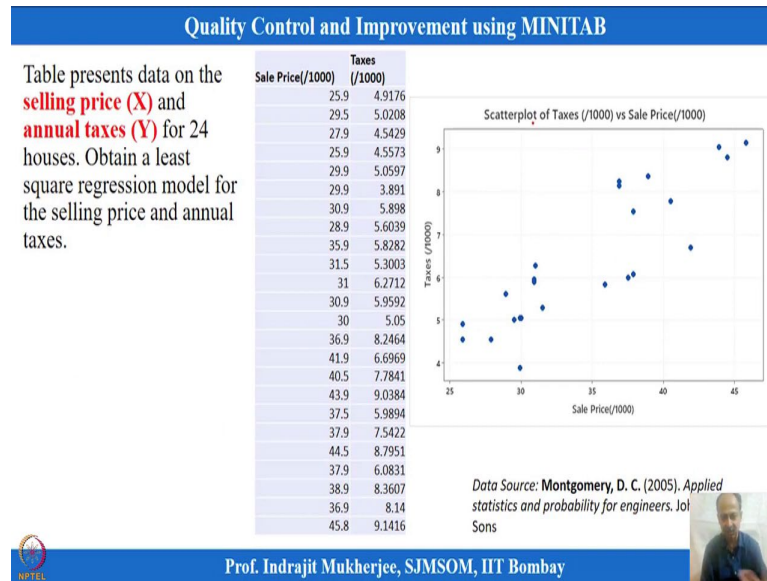
(Refer Slide Time: 26:01)



So, this is the basic interpretation of this and we can generate the errors because regression equation we have got. So, actual value and predicted value will give me residuals and residuals are probability plots are done over here Anderson darling test is done over here and P-value is greater than 0.05. So, that indicates that this is more or less normal.

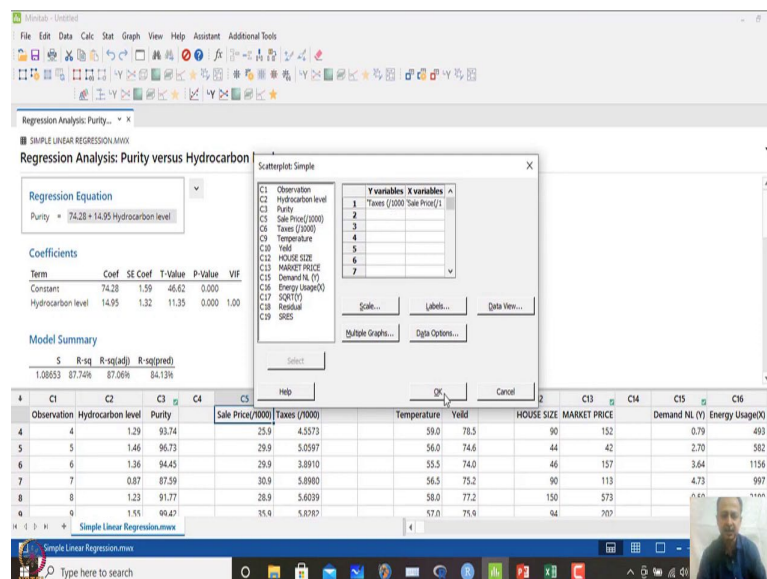
And the data set is on the line most of the points are on the line and also versus fit what do we see? There is no as such trend over here and more or less seems to be random. And also we can do the Breusch-Pagan test to confirm that one and other autocorrelation tests like Durbin Watson statistics can also be done.

(Refer Slide Time: 26:38)



This is another example of selling price and annual taxes I will just repeat the analysis over here, so that one more examples we are taking over here. So, let me go to that examples and try to figure out how we have done this one. So, just repetition one more examples where we want to see that whether annual tax is related with the selling price of the house like that ok.

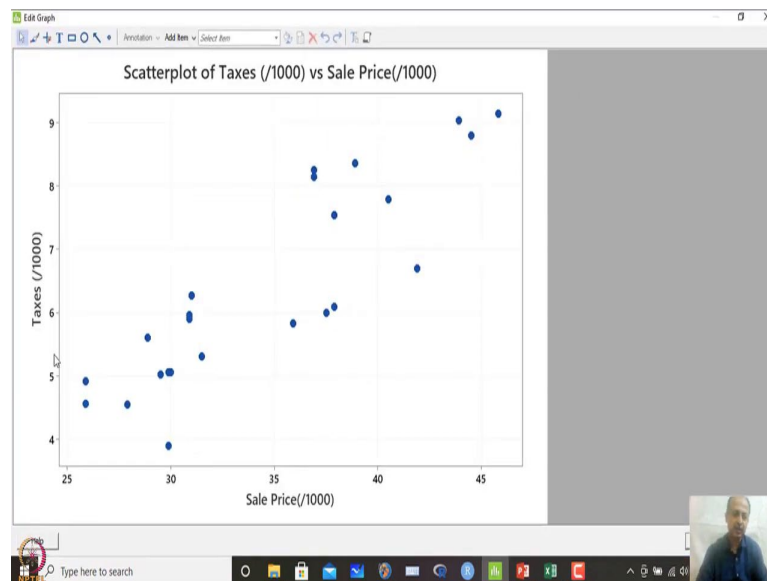
(Refer Slide Time: 27:10)



And we want to obtain the estimation relationship like that. So, what we do is that first step what we have to do is that this is the data set is given over here, this is the second

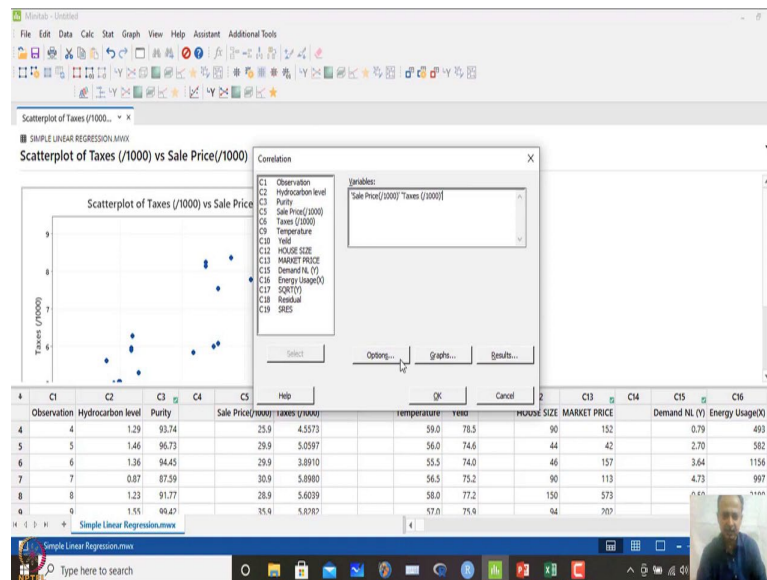
one c 5 and c 6 and this data is taken from again Montgomery's applied statistics and probability and I want to see the relationship of tax with sales price and tax will be y and sales price will be x over here. So, this is dependent on the sales price. So, what we will do is that we will we can plot the scatter diagram and try to see whether linear relationship is we can think of.

(Refer Slide Time: 27:37)

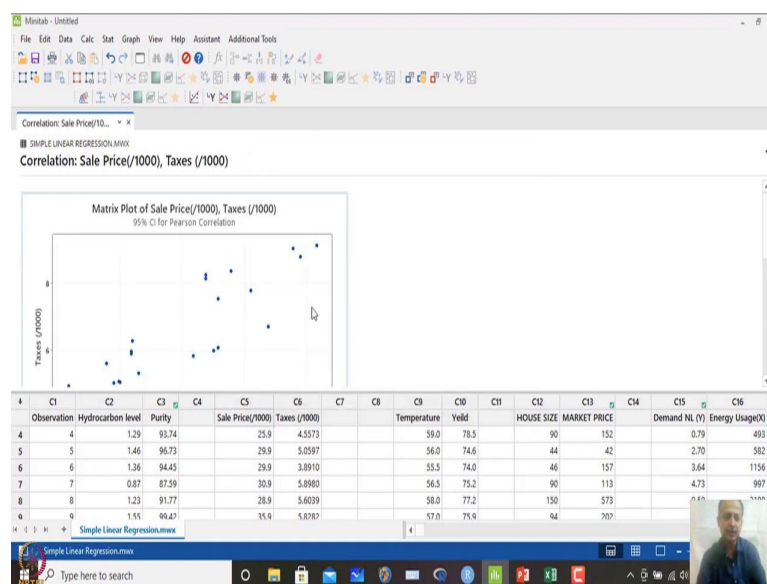


So, this y will be tax and x will be sales price and we can draw the graphical scatter plot over here and what you see in the scatter plot is more or less again positive relationship that exists over here and I can confirm what is the correlation coefficient between these 2.

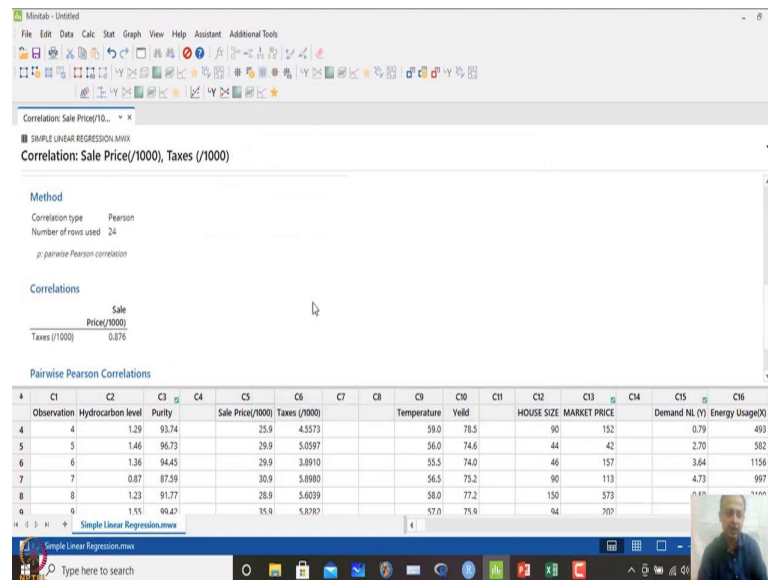
(Refer Slide Time: 27:48)



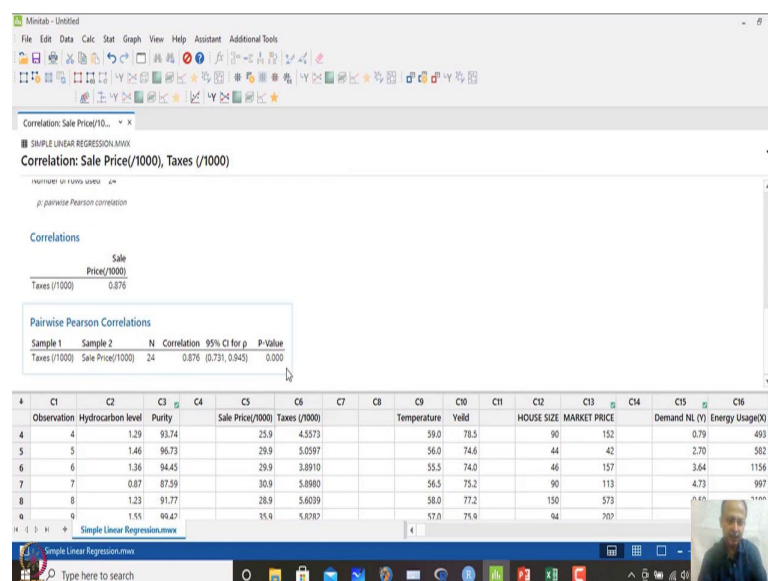
(Refer Slide Time: 27:52)



(Refer Slide Time: 27:52)



(Refer Slide Time: 27:54)



And I can go to correlation and I can just check sales price and taxes over here and I click ok and what I see is that P-value is significant over here and 0.876 is the more or less approximation that we are getting over here. Correlation coefficient this is positive; that means, there is a positive relationship between sales and tax that is relationship that we have.

(Refer Slide Time: 28:16)

Correlation: Sale Price/1000, Taxes (/1000)

Correlations

	Sale Price(/1000)
Taxes (/1000)	0.876

Pairwise Pearson Correlations

Sample 1	Sample 2	N	Correlation	95% CI for
Taxes (/1000)	Sale Price(/1000)	24	0.876	(0.731, 0.9)

Regression

Responses:

- Taxes (/1000)

Continuous predictors:

- Sale Price(/1000)

Model... Options... Coding... Stepwise...

Validation... Graphs... Results... Storage...

Help

OK Cancel

C1	C2	C3	C4	C5	C6
Observation	Hydrocarbon level	Purity	Sale Price (/1000)	MARKET PRICE	Demand NL (Y)
4	1.29	93.74	29.9	5.0597	56.0
5	1.46	96.73	29.9	3.8910	55.5
6	1.36	94.45	30.9	5.8980	56.5
7	0.87	87.59	28.9	5.6039	58.0
8	1.23	91.77	31.4	5.8762	57.0
9	1.51	90.42			75.9

(Refer Slide Time: 28:25)

Regression Storage

☒ Residuals

☒ Standardized residuals

☐ Deleted residuals

☐ Leverages

☐ Cook's distance

☐ DFFITS

☐ Coefficients

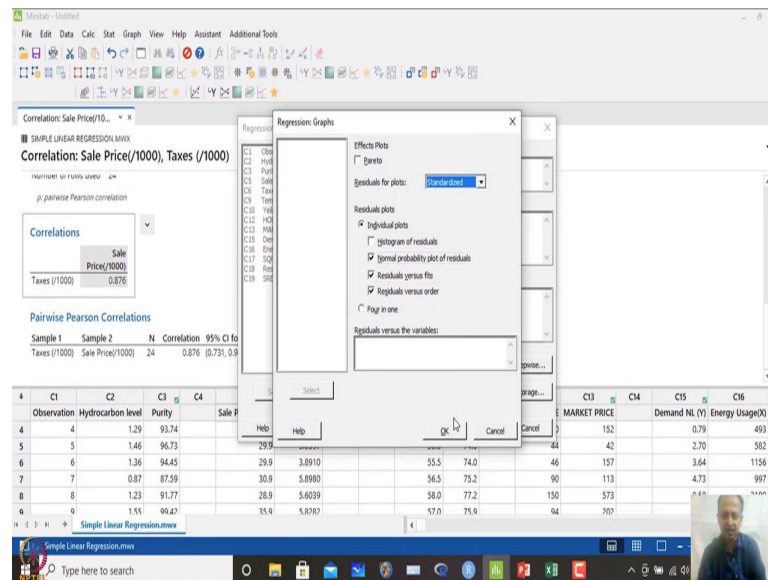
☐ Design matrix

Help

OK Cancel

Storage...

(Refer Slide Time: 28:31)

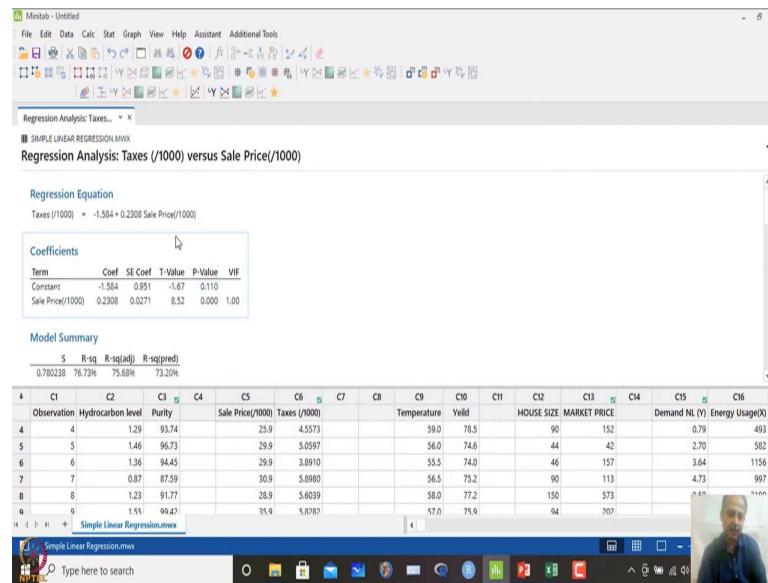


Now, everything is fine so I want to predict the equation. So, I will go to regression and regression over here fit regression models over here and instead of this purity we will change this one to tax and instead of continuous variable I will sales purity I will use over here.

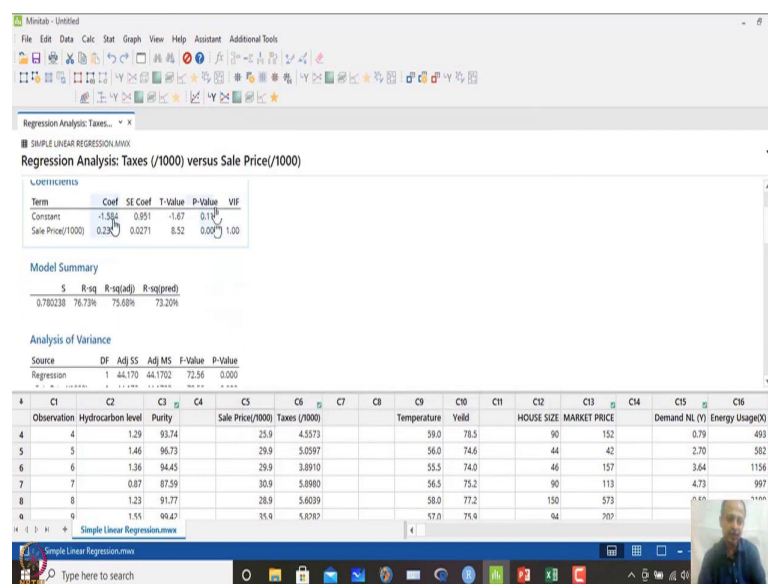
And I can store the residual again over here standardized residual and I click that will be saved at the end and then I will click on the graph. So, I can see residual plots I can also see normal plot of the residual and residual for plot may be standardized residual I want to check and then I will do ok.



(Refer Slide Time: 28:41)



(Refer Slide Time: 28:44)



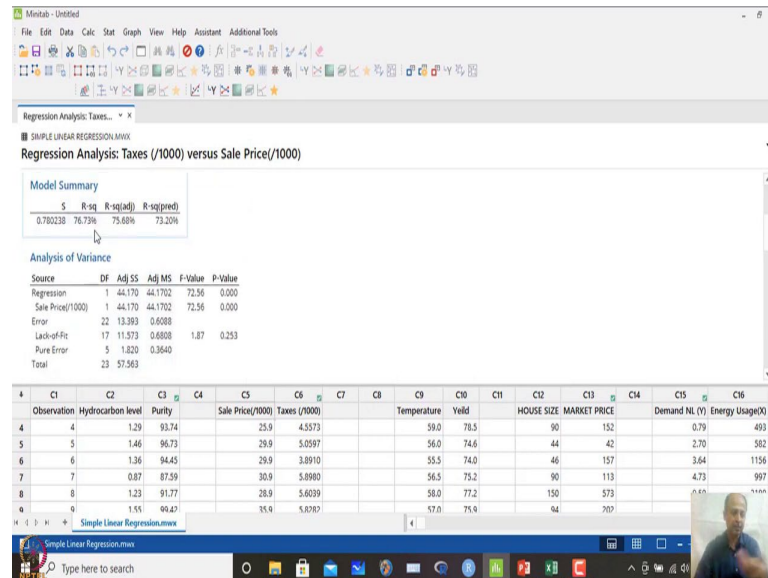
And then what will happen is that it will give me the equations, then it will say whether beta is constant is significant or not. So, what we see is that constant is not significant, but I told that researcher suggest that we should keep the constant. So, we will keep that one and sales price is significant over here; that means  $\beta_1$  is significant. So, we will retain this one, so there is a positive relationship that we are seeing.

So, coefficient is positive over here ok and constant is negative that constant cannot be interpreted in regression. So, in that case physical interpretation is not possible. So, and



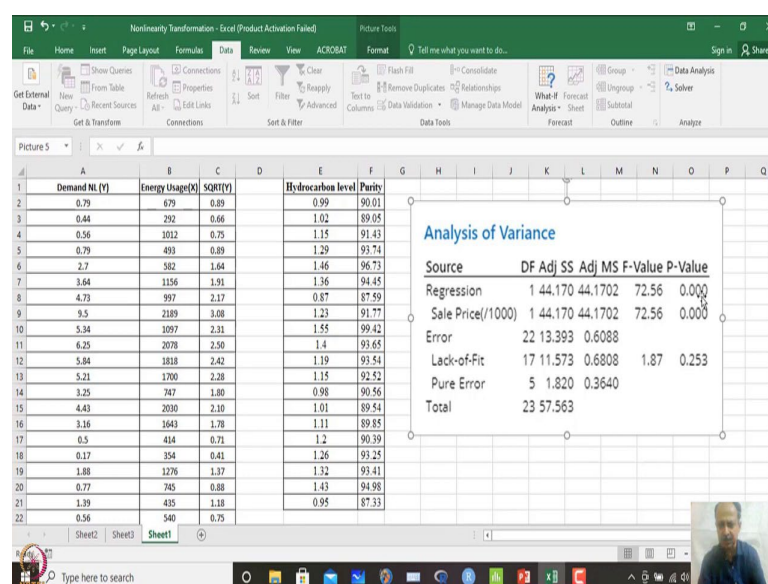
the R square value is 0.7673 what we have seen calculation also previous calculations like that.

(Refer Slide Time: 29:19)



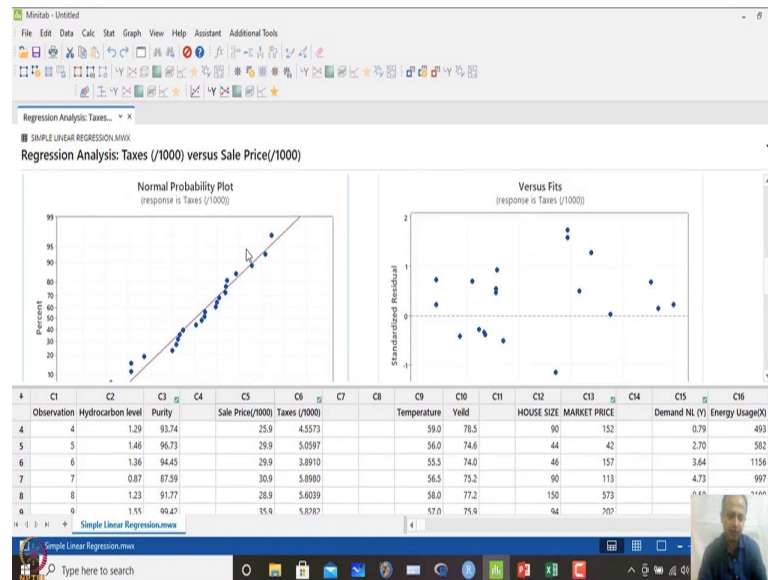
So, it will be close to R square that is calculated for manual analysis regression by its just a total and it is significant and there is no lack of fit; that means, 0.25 that you are seeing over here. So, approximately what we can see is that copy if I copy this image and I paste it over here in another sheet.

(Refer Slide Time: 29:35)

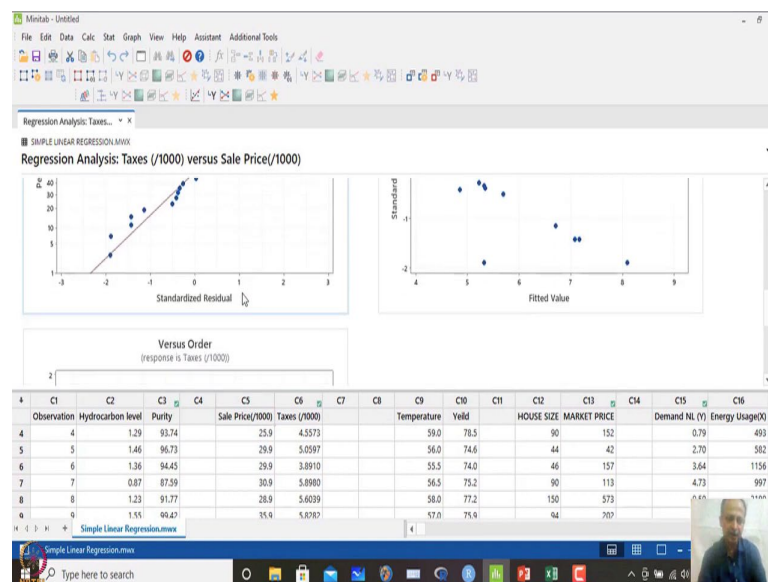


So, what we are seeing over here is that analysis. So, in this case what we are seeing is that regression is significant black of it is not there. So, in this case model can be generalized.

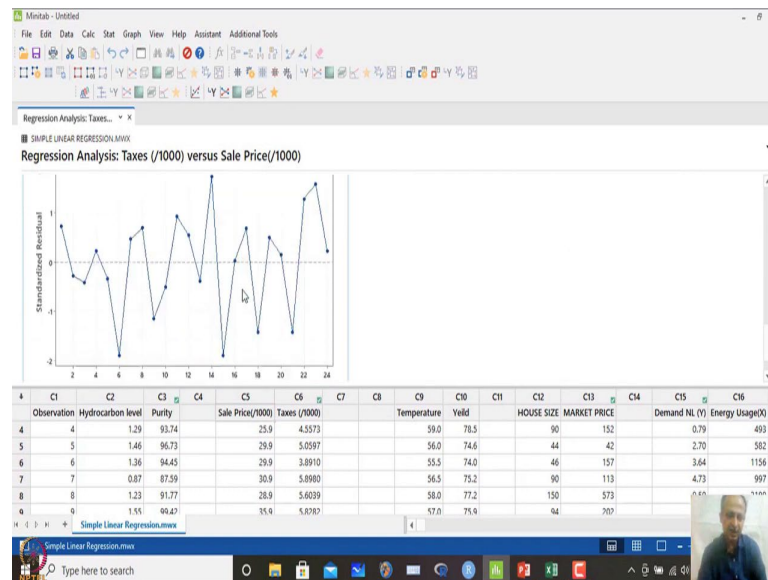
(Refer Slide Time: 29:47)



(Refer Slide Time: 29:49)

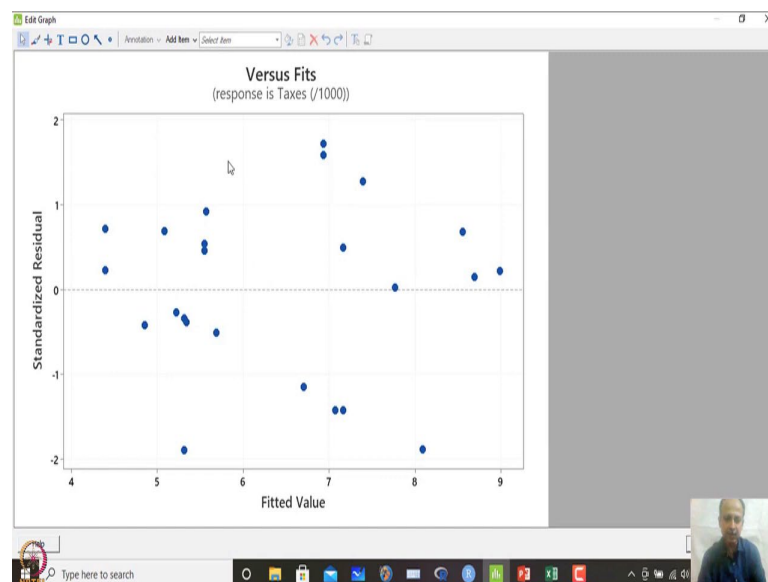


(Refer Slide Time: 29:51)

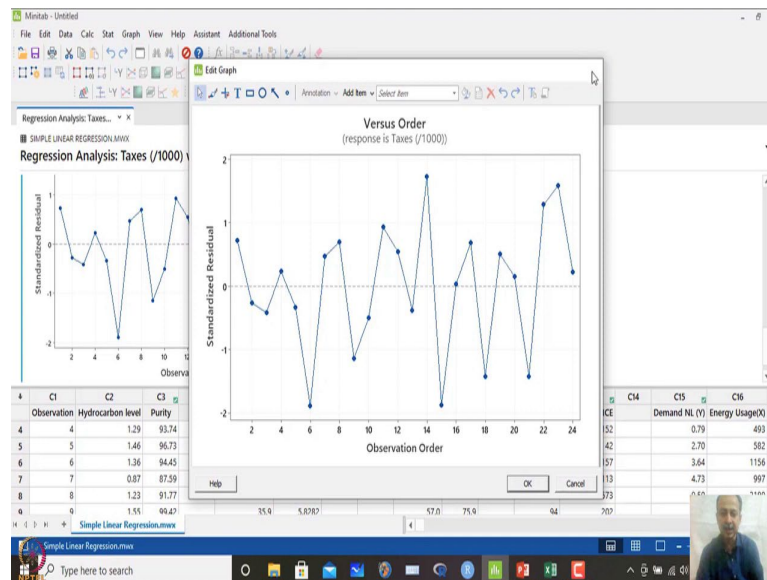


So, normal probability part more or less is seems to be ok, but we can do Anderson darling test and there is as such no pattern in the residual versus fit diagram equation over here.

(Refer Slide Time: 29:54)

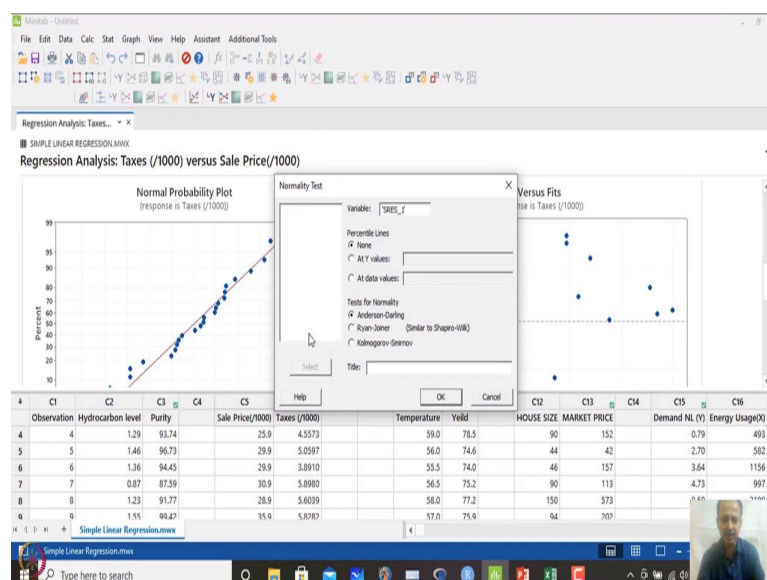


(Refer Slide Time: 30:04)

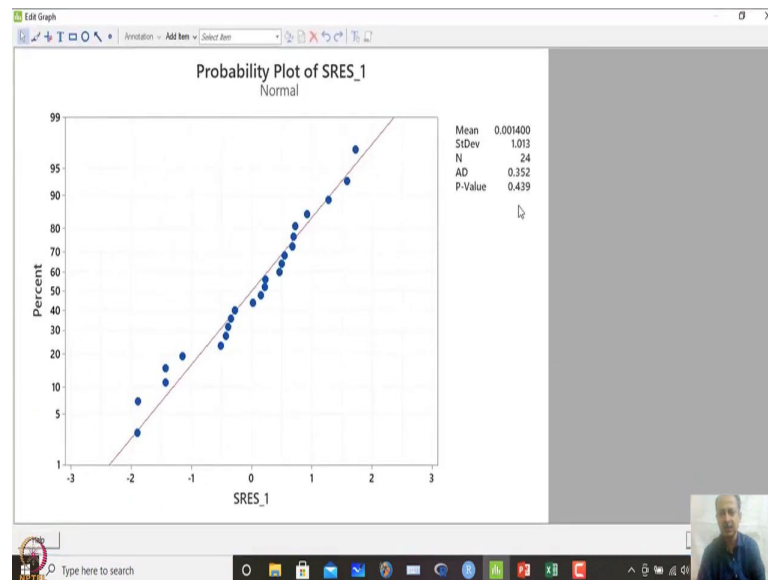


So, heteroscedasticity may not be a problem over here and also the autocorrelation does not seem to be significant, because there is no trend as such that we are seeing and maybe random observations. But we can do Durbin Watson we can do that test Durbin-Watson test and or Breusch-Pagan test BP test we can do over here to confirm that one ok and we have stored the residuals.

(Refer Slide Time: 30:23)



(Refer Slide Time: 30:26)



So, we can see at least the normality test and when I go to the last residual over here and do this test what will be what we see is that the p-value is not significant. So, in this case we can assume that the residual is normal. So, what we do is that we test the residual over here and if the condition is not satisfactory then in that case we go for transformation.

So, next case what we will see is that ok errors are non normal. So, in that scenario what is to be done and how the regression models has to be developed. So, that we can see in another examples in our next session like that and some more complexities we can see for now we will go to multiple regression.

So, some time we will spend in multiple regression, because in design of experiment what has what happens is that there is not single x that we are dealing with design of experiments. There can be multiple x single x is the one way analysis of variance that we are dealt with and that is the simplest of condition but that is not reality basically.

So, there will be multiple x that will influence the y and we need to develop the mathematical model and we need to develop the regression equation which we have to optimize basically finally ok. So that case we will discuss about that, but at present we will stop over here and we will continue in our next session on simple regression complexities when error is not normal in that case how do I deal with that.

Like analysis of variance and we will also go to multiple regression and see the complexities what, in modeling what we face basically. So we will continue with that.

Thank you.