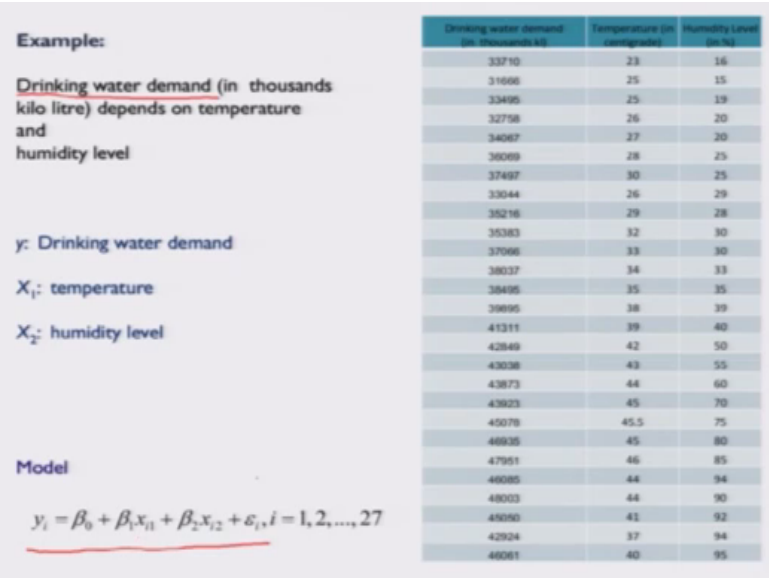
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture 19
Software Implementation of Multiple Linear Regression Model using MINITAB

Welcome to the lecture you may recall that in the earlier lectures we had considered the set up of multiple linear regression model and we discussed different types of aspects related to the fitting of model judging the goodness of fit of a model and finally the objective was to get a good fitted model. Now let us take a simple example to understand whatever we have done in the earlier lectures, right. So if you see here am going to consider here an example where I am considering 27 observations.

(Refer Slide Time: 00:57)

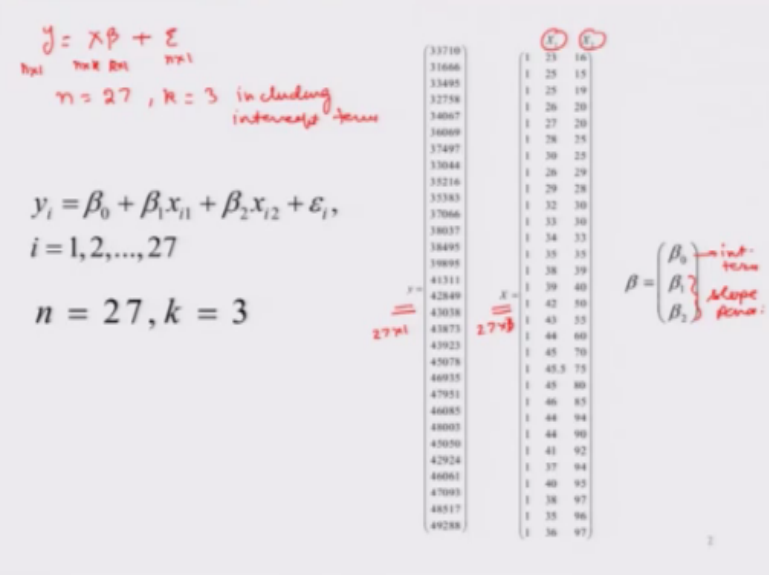


Drinking water demand (in thousands kl)	Temperature (in centigrade)	Humidity Level (in %)
33710	23	16
31666	25	15
33495	25	19
32758	26	20
34067	27	20
36069	28	25
37497	30	25
33044	26	29
35216	29	28
35383	32	30
37066	33	30
38037	34	33
38495	35	35
39895	38	39
41311	39	40
42849	42	50
43038	43	55
43873	44	60
43923	45	70
45078	45.5	75
46935	45	80
47951	46	85
46085	44	94
48003	44	90
45050	41	92
42924	37	94
46061	40	95

And these observation have been obtained on the demand of drinking water, which is measured in 1000 kilo liter and we believe that this demand depends on the temperature that is the weather temperature and humidity level. You may also recall that in the case of linear regression model we also considered a similar example in which I have taken only variable that is temperature.

And now I am going to consider the second variable as humidity level and that is measured in percentage. So the model in which we are interested now here is y=0 + beta1 x1 + beta2 x 2 + epsilon and we had got 27 observation so I can express the model as yi= beta0+beta1 xi 1+ beta 2 xi 2+ epsilon i.

(Refer Slide Time: 02:08)



Now first let me explain how these things can be expressed in the framework of a multiple linear regression model. So here the model which we had considered was expressed as y= x beta+epsilon where y was a n cross one vector of observations on response variables and x was a matrix of order n by k of k independent variable having n observation on each of the variable and beta was a k cross one vector of regression coefficients and epsilon was a n cross one vector of random error components.

So now in this case you can see that here n=27 and k is equal to here actually three including the intercept term, okay, so now I can express all the 27 observations on y here like this which is a 27 cross one vector and there are two variables x1 and x2 on which we have got 27 observations and all these observations including the intercept term can be expressed as a matrix x which is of order 20 by 3.

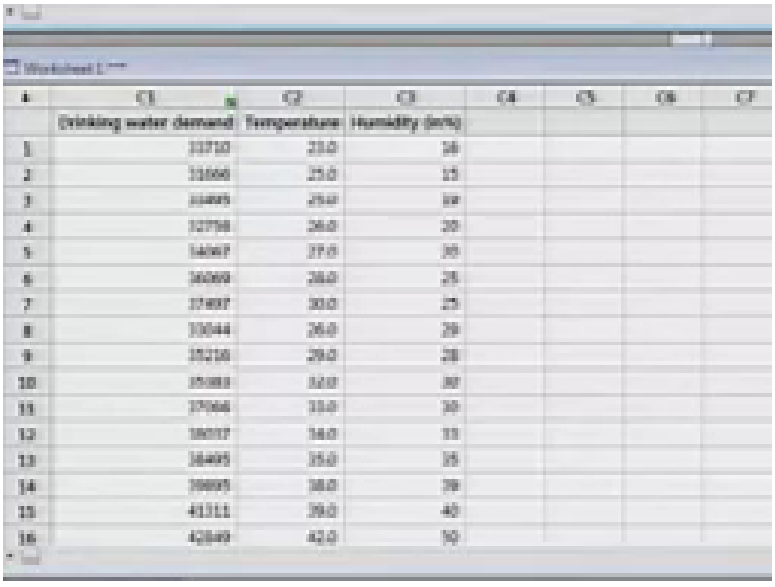
And similarly beta contains three elements first is intercept term and other two are slope parameters, beta1 is the slope parameters associated with x1 and beta2 is the slope parameter associated with x2. Now let us try to use software which can analyze this data, so as discussed earlier we had seen that there are different types of statistical software available and some are paid software and some are free software among the paid software Minitab, Systat, SPSS, SAS and all such softwares are quite popular.

Among the free software R or GRETL these are two popular softwares among others, right. So in this lecture we are going to use the software Minitab, but definitely if you use any other software that will also give you the same outcome there can be some difference in the format that some software may give ANOVA table first and then the regression coefficient and some software may give the regression coefficient first and then the ANOVA table and so on

So there are only minor difference among the output and outcome of software, so that doesn't make any difference. Now before we go forward let me explain you what is our objective, my objective is not to find a model here by objective is that first we to expose this dataset to software and whatever is the outcome of software we have to understand that thing. So the first question is this, what different numerical values are indicating and how they have been computed and they correspond to wish of the quantity that we have covered in the lecture.

And after that we will try to analyze all the things one by one and finally we will try to combine all our outcomes together and then we will have a final fitted multiple linear regression model, right. Let us now start with this thing and I already have entered the data in this Minitab software.

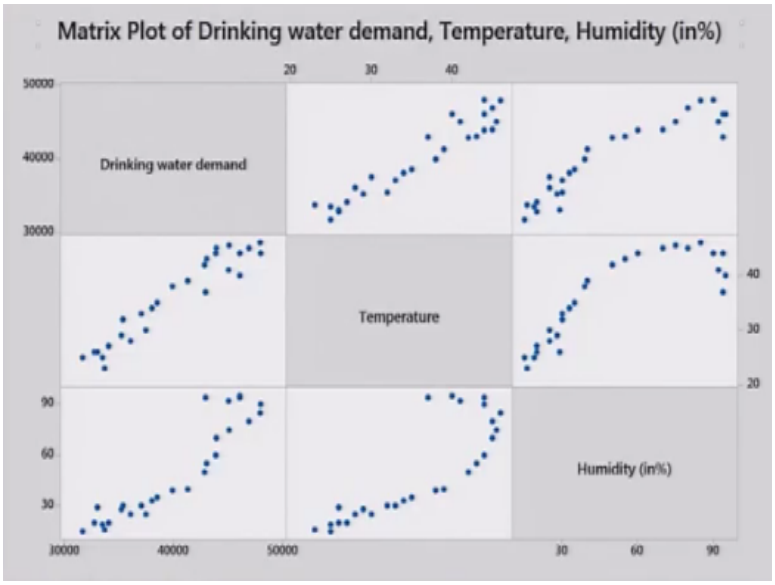
(Refer Slide Time: 06:00)



So you can see here that in the first column I have entered the data of drinking water demand the second column has the temperature and third column has the humidity. Now first of all we have to ensure that can we really use here multiple linear regression model, so we try to plot them since this is a multiple linear model so we would like to go for multiply scatter diagram or in the case of Minitab this is called as matrix plot.

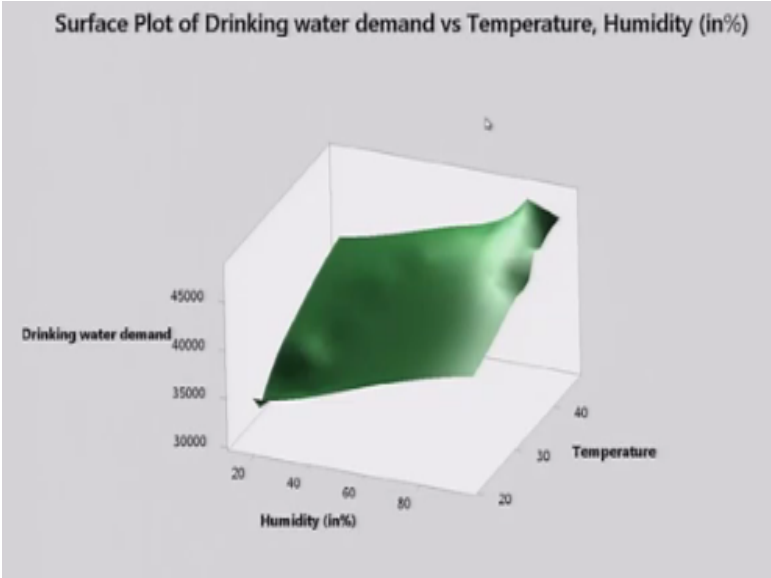
Right, so now there are different options available for the matrix plot and then we are just going to opt for a simple matrix plot and we can see that here I have chosen here all the three variables c1, c2, c3 and once I try to plot it I get here a scatter diagram like this one. Okay, let us try to first live it a side and try to make a some other things and at the end I will try to copy all the commands and then we will try to discuss them one by one.

(Refer Slide Time: 06:52)



Now since we have here only two variables x1 and x2 and third variable is here y, so one option we have that we can also make a three dimensional plot and so we try to go for a three dimensional plot, which is based on the wire. So here I try to using the Z axis is try to use the drinking water variable on y temperature and on X axis the humidity and we get here this type of three D surface.

(Refer Slide Time: 07:42)



Now let us try to store it and we will try to analyze it later, and similarly there is another option here that we can have here a three D surface plots, if try get the surface plots here that will give us this type of surface. So I am just trying to prepare the ground and we would like to be confident that once I fit a linear regression model we should be confident, yes this is situation where a multiple linear regression can be fitted.

Obviously when have here more than three variables all together than the option of three dimensional surface plot is not available and then in that case we have different options that we can take say three variables at a timed and then we try to combine all the individual inferences together and finally the objective is to take one regression in terms of obvious or no that can we fit here a multiple linear regression model or not.

Right, so these figures are giving us a sort of confidence that, yes a multiple linear regression model can be fitted over here. So now let me try to fit here a multiple linear regression model so you can see that in Minitab it is not difficult and it is a menu driven software. So now here I am taking my response variable as a drinking water demand and I have taken here two independent variable as temperature and humidity and here we do naught have any categorical variable which takes the values in terms of say zero one or something like that.

Let me try to explain what are the different meanings which are available to us, so here you can see that there is possibility here that one can also at the interaction term in multiple linear regression model once we decide yes two variable are interacting then the observation on

interaction are difficult to find in practice so what we try to do we simply try to multiply the values of the two corresponding variable and we obtain the observations on interaction.

But here in this case we are not considering the interaction term so we ignore it. Now if you see here we have here different options now I am setting here my this confidence level for all intervals as 95 that means the level of significant alpha is here 5% or point 05 now when am going for confidence interval I can have one sided interval are two sided interval so we decide in the case of in the of linear regression analysis we are usually interested in a two sided confidence intervals, so I give this option over here.

Well, now when we want o find out the sum a square there are two options that I can find out the adjusted sum of a square or say sequential sum of squares, but in the case of linear regression analysis what we have considered we are interested in the simple sum of a square are that is called here as a adjusted sum of a squares, right adjusted had some other meaning but that is beyond the scope of this lecture.

Another option here is this Box-Cox transformation that we have not considered in this lecture, but this is a sort of transformation that is applied to the independent variables if y is not really related to x, but it is linearly related to some other power of x something like square root of x or x square something like this, but we are not considering here so we have just using to define the level of alpha and the type confidence interval and the type of sum of squares.

Since we are not using here the categorical variables, so coding is not important and we are not using here the stepwise regression, stepwise regression is another type of regression, which we have not described in this course. Now there is another option to have here the graphs, so you can see here that there are different options, histogram of residues, normal probability plot, residual versus plot, residual versus order.

And you may recall that when we discuss the diagnostic for our multiple linear regression model we had considered all the things but in this case I would be more interested in knowing whether the observations are originating form a normal distribution or not and then I would try to have a plot for the residual versus fits, so that I can decide for the presents of heteroskedasticity in my model.

Similarly for the residual versus order we have a time series plot so we can have all these things, now here if you see now we have here three options that we can make this residual plots with respect to simple residuals that we have defined by epsilon i hat are the standardized residual that we have define say as di or say ri say studentized residual or standardize residuals and deleted residual that we had defined as e and i inside the bracket.

So here we would like to have the regular one, and then we would also like to see that about what happens when we are going for a standardized one, ideally there should not be any difference. So in the first shot I will try to obtain the these plots for the regular residuals and then I will try to repeat this analysis and I will obtain the plots for the standardized residuals and there is another option that I can have all the four graphics in one but anyway that doesn't make any difference.

Now the another option is here results that what type of results I would like to have, one thing is the method, yes we would like to know that what type of method we have used like least square principal or something like this then we would also like to know what are the outcomes of analysis of variance and we would also like to know the model summary that consist of fourth the statistics like r square adjusted r a square PRESS and so on and we would also like to store the values of a coefficients that is the values of beta zero hat, beta1 hat and beta2 hat, and we would also like to have a regression equation that is obtained after fitting the model.

And we would also like to know the fits and diagnostic if you remember we had done different types residuals and different types of statistic like as hii is the ith diagonal element of hat matrix, deleted residuals DFFITS and so on. Now I have here two options that whether that I would like to fits and diagnostic for all the observation and or say only for those observation which are found to be unusual.

But since here we are learning, so we would like to see what happens to all the observation and then there is a statistics what is here called as a Durbin-Watson statistics we have not done this thing in detail but still I would try to its outcome and then I will try to explain you how to interpret it. Durbin-Watson test is a test to check the presence of first order

autocorrelation in the data that means incase if the data are affected by the time then we have time series analysis.

And in the case of time series analysis we try to test whether the observations have first order autocorrelation are something else. Right, now in case if you want you can also store these results whatever you have obtained here, right so this can be stored in the worksheet that is opened here and one may use it. So we would try to store to just illustrate that how this actually works.

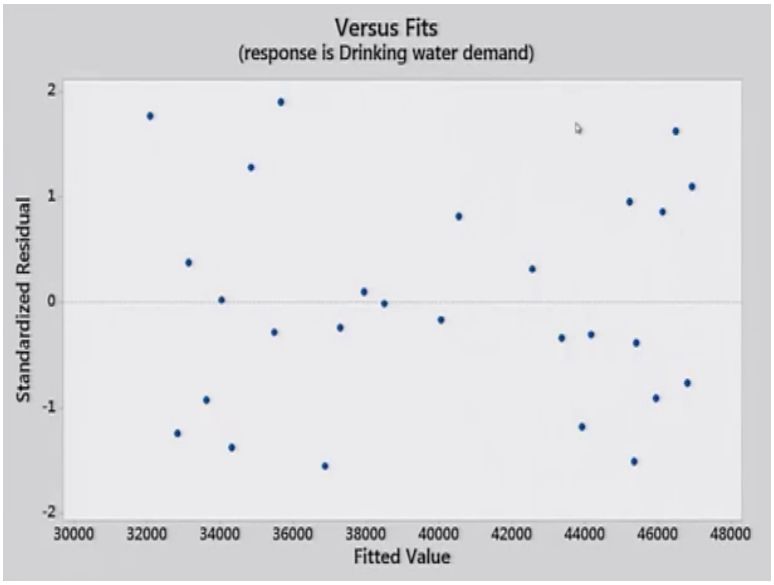
So now once I go for this analysis let us try to see what happens, so now you can see here that we had obtained here a normal probability plot so we try to store it and we have got here a plot between residuals and fitted value, okay. So let us try store it, now if you observe over here this is the outcome here we have analysis of various results in here and here we a have the details about the model summary which is consisting about r squares, r square adjusted PRESS and r square prediction.

And similarly here we are getting the value of the regression coefficient their confidence interval various inflation factor values for the t statistics and here we are getting the fitted regression equation, and here there are different values of diagnostics and since we had given that we would like to have this diagnostic for all the observation so this is giving here say the fit standard error of it confidence interval for the y hat residual, standardized residual, deleted residual hi and Cook's D statistics here you can see.

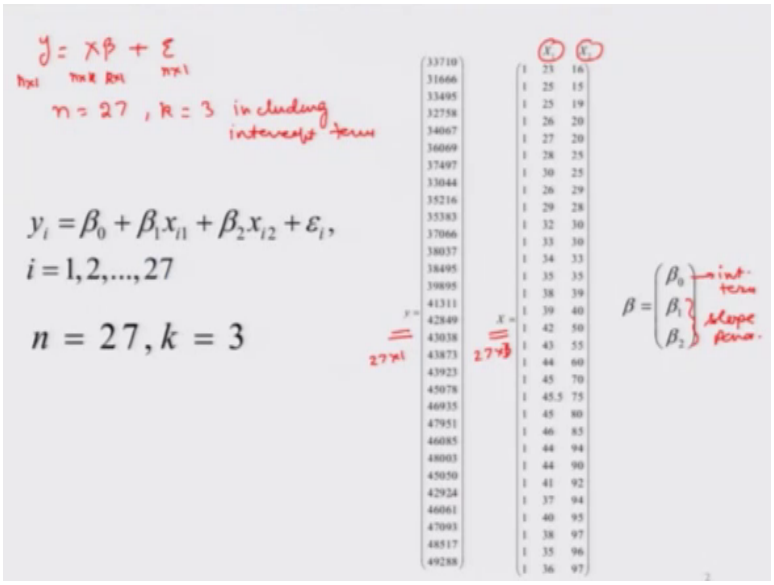
Right, and this is available for all the 27 observations and similarly the DFFITS statistics is also computed here for all the observation and at end we have here Durbin-Watson statistics and we try store these results separately and then we try to analyze it, but before go for the further analysis let us also try to obtain the this residual plots for the standardized case.

So now I can choose here this option standardized and let me do this analysis, so you can see here that we had now here these two graphics when the residuals are standardized this is the normal probability plot and this here the standardized residual versus fitted value plot.

(Refer Slide Time: 18:58)

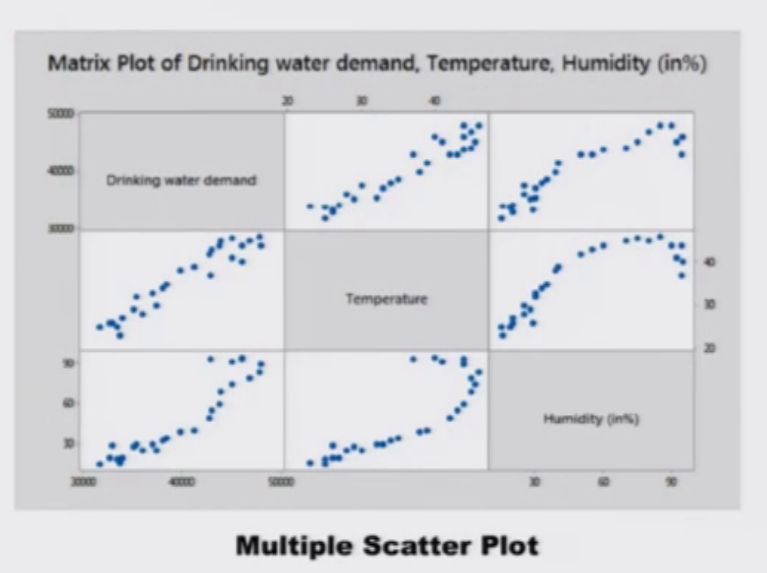


(Refer Slide Time: 19:04)



So now we have stored all this information over here and we would try to understand the interpretation and other things for these results.

(Refer Slide Time: 19:15)

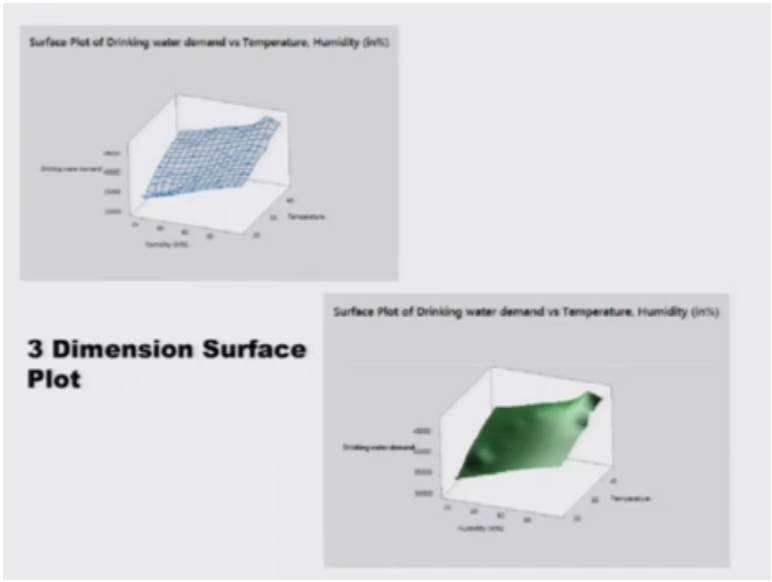


Multiple Scatter Plot

So you see this is the matrix plot of the three variables y x1 and x2. One can see here this is a plot between y and x one and one can see here that it is showing as a linear trend so there is no issue similarly this is a plot between y and x2 and this is also showing here a linear trend and similarly this is also a plot between x one and y this is showing a linear trend and similarly this is also a plot between x2 and y this is showing us a linear trend.

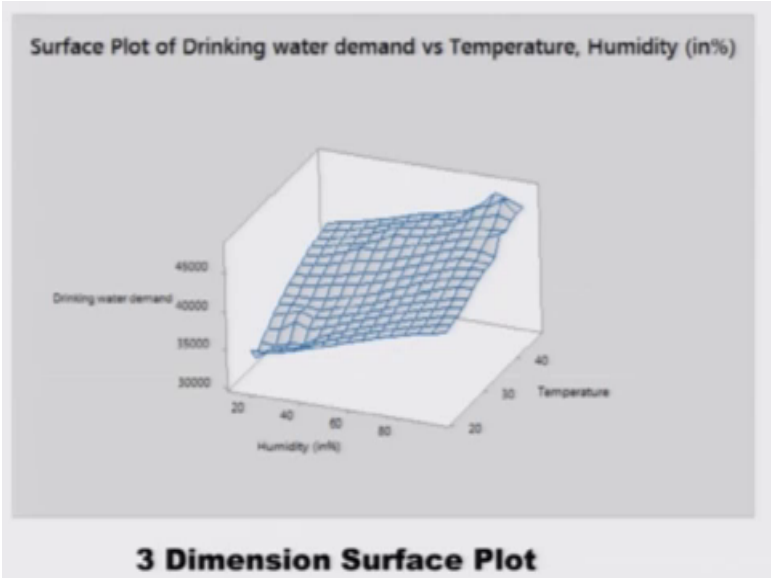
Now if you try see here this is a plot between x1 and x2, you can see here ideally we assume that x1 and x2 should be independent so we should have a scatter diagram like this one, but here you could see it is trying to show here a sort of linear relationship. So let us see whether this is reflected in our regression analysis or not, but more or less this is giving us a surety that at least the relationship between y and x one is linear, relationship between y and x2 is linear.So we can believe that the relationship between y and x 1 x jointly will also be linear.

(Refer Slide Time: 20:43)

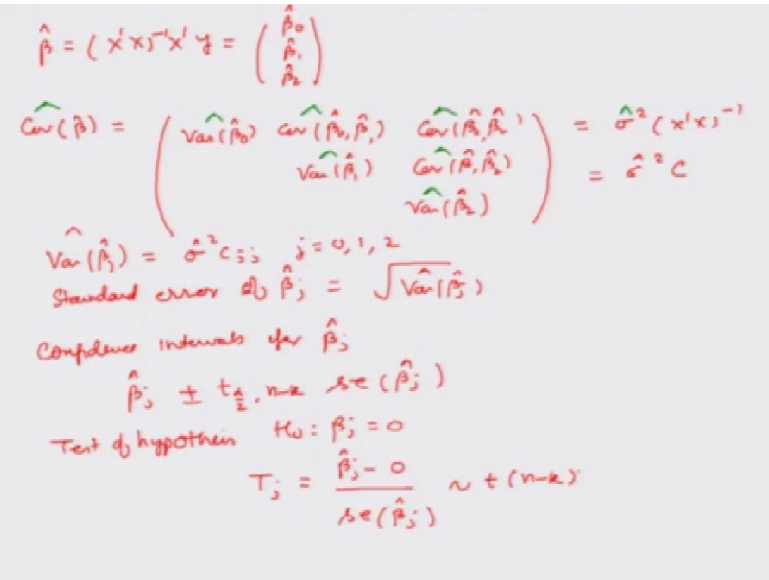


Now these are the surface plots that we had obtained you can see here that in deep case this is also showing us sort of increasing a trend and this is the same story her also this also showing a sort o f increasing trend so these three dimensional plot which are useful particular in this example are also giving us confidence that a multiple linear regression model can be fitted in this case, right okay so this is the same thing that we had obtained it is in a bigger shape.

(Refer Slide Time: 21:08)

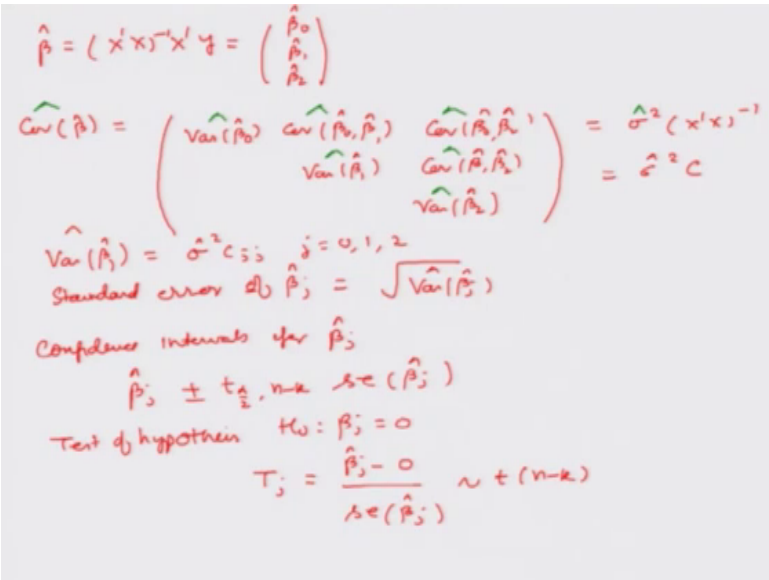


(Refer Slide Time: 21:26)



Now after this we are going to discuss about the outcomes of regression analysis, but before that let me briefly describe some results so that it is easier for us to understand you may recall that we had estimated beta hat say x transpose x whole invest x transpose y.

(Refer Slide Time: 21:30)



So in this case this is going to be like beta0 hat beta1 hat and beta2 hat, and a three cross one vector of these three values and similarly we have obtained the covariance matrix of beta hat, and which in this case will variance of beta0 beta naught variance of beta1 hat and variance of beta2 hat, and on the off-diagonal elements this will be covariance between beta0 hat beta1 hat and covariance between beta0 hat and beta2 hat.

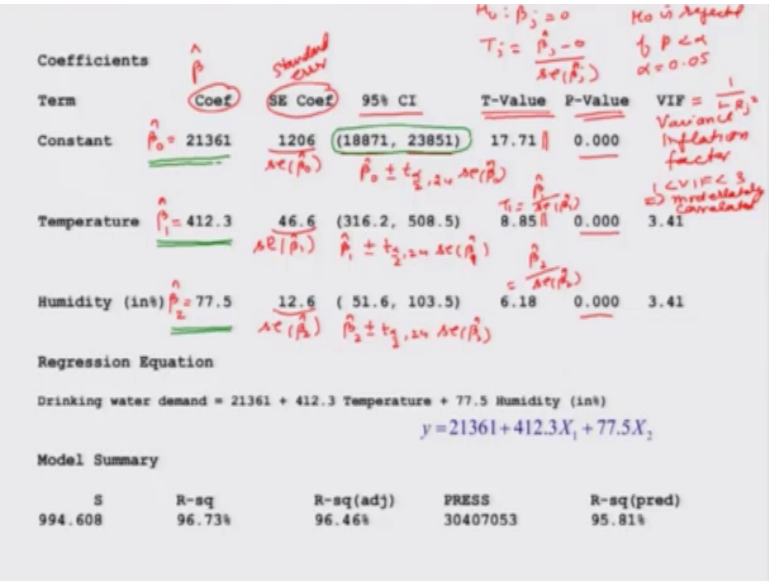
And here the covariance between beta1 hat and beta2 hat so this matrix will look like this and but our interest here is to go for its estimate, so the estimates and this value was obtained as

sigma square x transpose x whole inverse but here we are interested in the estimate of covariance matrix, so that we had obtained like this so this will be the estimated values and this will sigma square hat x transpose x.

Okay, and you may recall that we had denoted for example this sigma square some matrix here see c where we see that variance of beta j and its estimator is given by sigma square hat cjj, j goes around here in case of zero1 and 2. So that is the symbol that we are going to use here and when we are trying to find out the standard error then standard error is of beta hat j this is simply the positive e square root of estimate of variance of beta hat j.

Then we also had obtained the confidence interval, the confidence interval were obtained as say for say beta hat j this was obtained as beta hat j+- t alpha by 2 and - k and standard error of beta hat j and for the test of hypothesis when we are interested in testing the hypothesis h naught beta j=0 then the t statistics for the jth regression coefficient was obtained as beta h j -0 divided by standard error of beta hat j, and this was following a t distribution with n-00 k degrees of freedom.

(Refer Slide Time: 24:39)



Now you will see that the values which we have obtained earlier analytically now they are being computed easily by the software so let us try to understand here the this outcome I have divided the entire outcome into separate sheets for that we can understand them clearly. Let us go one by one, so when we are trying to see here this coefficient, coefficient is nothing but your beta hat vector so this is here beta hat zero this is the constant.

Constant means the estimator of intercept term and the temperature means here the regression coefficient corresponding to the independent variable temperature x1 and similarly this is here the value of beta two hat, which is associated with the second independent variable humidity. Now this is the standard error, so these values are giving us the idea of standard error of different coefficient, so this is standard error of beta hat naught, this the standard error beta1 naught and this is the standard error beta hat 2.

The next column is about the 95% confidence interval, so this is the lower confidence limit and this is the upper confidence limit for beta hat naught this is provided by this expression t say alpha by 2 and with degrees of freedom n - k, which is 27-324 and standard error of beta hat naught and similarly this is also a confidence interval foe beta1 hat obtained like where alpha = 05.

And this is the confidence interval for beta hat 2 this is obtained at alpha = point 05 with 24.degrees of freedom and standard error beta hat. So these things give us this idea, so before we try to go for the next column let us try to understand what is the interpretation, the interpretation is that on the basis of given set of data we can say that the value of intercept term here is 2 1 3 6 1this is the point estimate.

And the confidence interval for beta naught is like this, this means the value of beta0 is going to lie between 1 8 8 7 1 and 2 3 8 5 1, so you can see that the point estimate is lying within the confidence interval and similarly the value of this beta1 hat is indicating that the point estimate is 4 1 2 . 3 where as the confidence interval is 3 1 6 . 2 2 5 0 8 . 5.

So one can also see here that this is going to lie between the confidence interval and the same thing is for this beta two hat also, right. The physical interpretation of beta hat naught is the following; that when x1=0 and x2 is also =0 then the average value of y is going to be 2 1 3 6 1 right, and similarly the physical interpretation of beta one hat is that when there is a unit change in the value of x1.

For example in this case x1 is that temperature, so when that the temperature changes by 1 degree centigrade then the change in the average value of y is 4 1 2 . 3 that means once the temperature changes by 1 degree centigrade then the demand for water on a average increases

by 412.3000 kilo liter the question is why I am saying it increasing because one can see here that the sign of beta one hat is positive.

In case if the sign would had been negative then I would say that the demand would decrease by 4 1 2 point 3000 kilo liter when the temperature changes by one degree. Okay, similarly here also in the case of humidity also it is trying to say that when there is a change of 1% in the humidity it increases the n the average value of y changes by 77.5 units.

That is when there is a 1% change in the humidity on an average the demand of drinking water in cases by 77.5000seven kilo liter, right and these are the indicators like as standard errors, this is giving us an idea about the spread of beta0 hat beta1 hat and beta2 hat, okay.Now we come to this column t value, t value is about the test of hypothesis that h naught beta j=0.

So for example this and here t j= beta hat j-0 divided by standard error of beta hat j, so this value is beta hat naught divided by standard error of beta hat naught this is the value of t statistics for the intercept term and similarly this is also the value of beta1 hat divided by standard error of beta1 hat that we had denoted by here as t one and similarly this is also a value which is beta hat 2-0 divided by standard error of beta2 hat, so these are the values of t statistics.

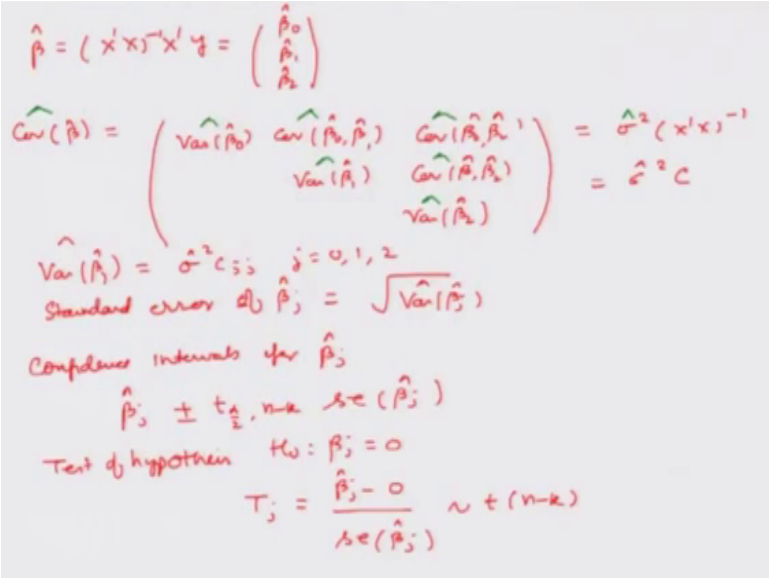
Now as I told you earlier that in order to test the hypothesis we have two options either we can obtain the critical values or the tabulated values from the probabilities of t tables or in software there is another option to use the p values and we had learned that the result is that h naught is rejected if p value is less than alpha that is the level of significance so here we have taken alpha = 0.05.

So you can see here that the p value in all the three cases they are very, very close to 0, so one can say that in this case all the three null hypothesis that is h naught beta naught =0 h naught beta1=0 and h naught beta2=0 all are rejected. So that means none of the values beta0,beta1 or beta2 is 0 or close 2 0. This means the variables which we have considered as a temperature and humidity.

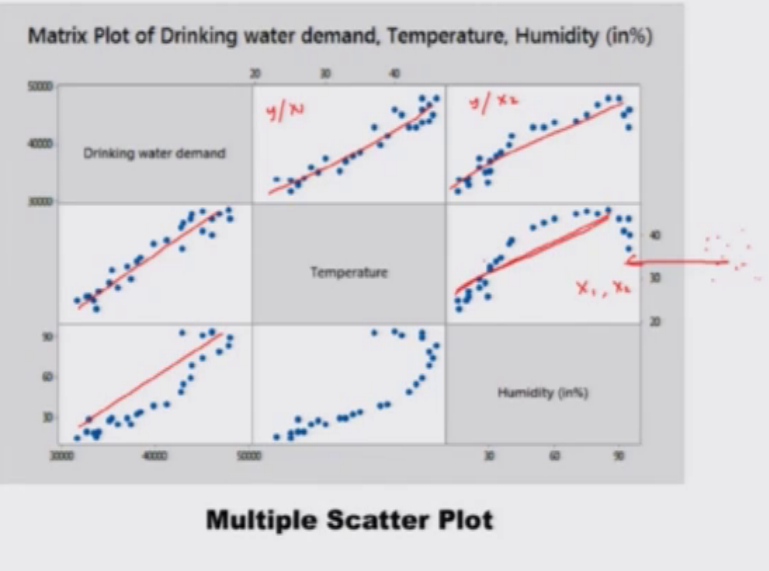
They are contributing in the model and they are helping in explaining the variation in the value of y. The last column is various inflation factor you may recall that we had discussed this diagnostic when we wanted to test the assumption of rank of x = k that means all x1 x2 x k are independent or not and we had defined that VIF was =1 over 1 - r j square.

Where r j square is the coefficient of determination that is obtain by regressing that jth explanatory variables x j over the remaining k-1 explanatory variables, and if you remember we had given the criteria that in case if VIF is lying between one and three this will be indicating that the independent variables are moderately correlated and if it is more than three one can consider that they are highly correlated.

(Refer Slide Time: 33:24)



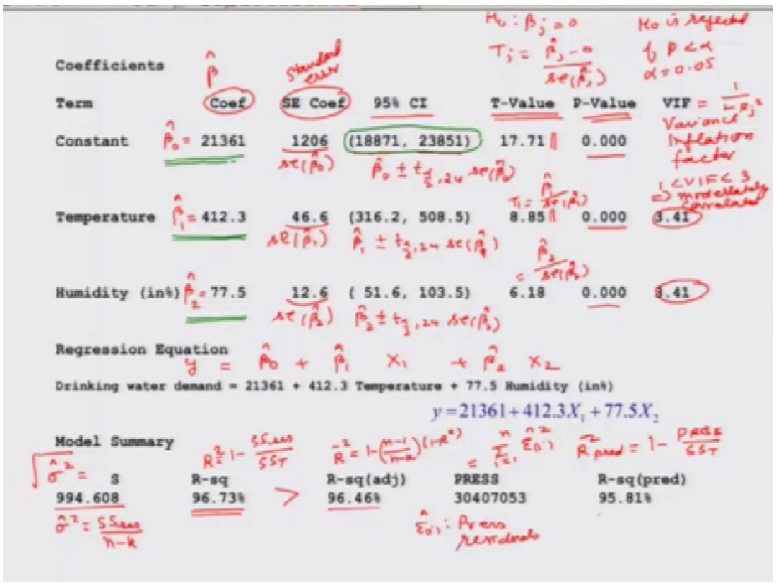
(Refer Slide Time: 33:43)



Multiple Scatter Plot

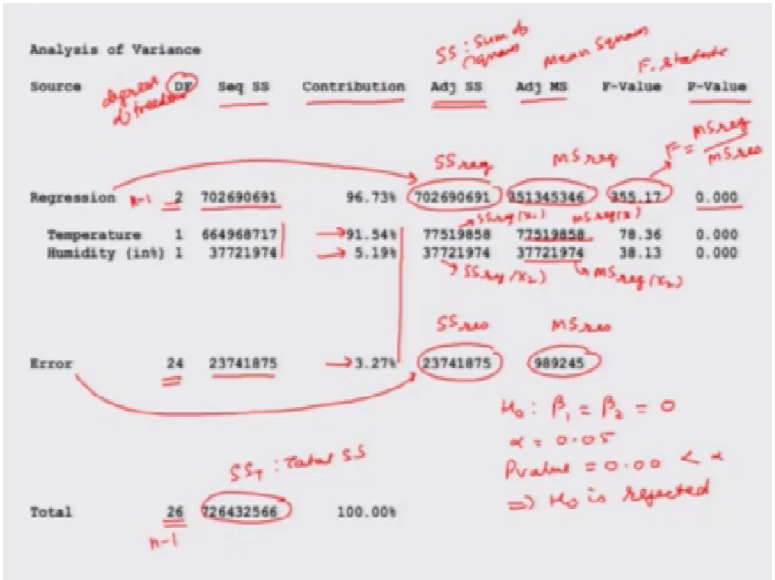
Now if you see here in this multiple scatter diagram which I indicated her you can see that that there is linear trend here a sort of linear trend so one can see here that x1 n and x2 are not really uncorrelated, but they have got certain degree of linear relationship and this is being indicated in this software here in this value 3.041 like this right okay. So you can see that our statistical analysis is also capturing this aspect.

(Refer Slide Time: 34:02)



Now based on this regression coefficients outcome we can now formulate our fitted regression model as y= this is the value of beta0 hat + this is value of beta1 hat this is x1 + beta2 hat times x two. Now after this we have here some more model summary this value here as S this is nothing but your square root of sigma square hat, and sigma square hat was obtain by SS residual divided by degrees of freedom n-k.

(Refer Slide Time: 37:31)



This is our here R square R square was a statistics that was used to test the goodness of fit and this we had defined as 1- SS residual divided by total sum of squares SST, so one can see here that this is 96.73, so one can be happier that the model which we have fitted is reasonably good and essentially the variables x1 and x2 are capable of explaining say at least 96.73% variation in the values of y.

Next quantity is here R square adjusted you may recall that we had define this quantity like R bar square and this was define as 1 over n-1 over n-k times 1- R square. So this is again say 96.46 which is again indicating that our model is nearly 96% good and you will always see that R square is always greater than the value of R square adjusted, but usually in practice this difference is very small.

Next here is PRESS, the PRESS was the statistics that was given by i goes from 1 to n epsilon i bracket whole square. You may recall that these epsilon i hats were PRESS residuals that means the different between yi and say yi hat which is obtain after deleting the ith observation and based on this PRESS statistics we had defined the R square prediction and which is obtain here by 1- PRESS divided by total sum of square SST.

And this R square predication was the quantity that was helping us in deciding whether the fitted model can be used for prediction or not and if used how it is going to perform. So this value here 95.81%, so it is simply indicating that this model is nearly 95% good for predication. So now let us try to concentrate on the on other part which is analysis of variance.

So here degrees of freedom means is denoted by here df, these are degrees freedom and here the software is computing other component sequential sum of squares and it is also computing the contribution of sequential sum of square, which we have not covered in our lecture, but I will try to explain you its importance over here, what we considered in our lecture is that we have obtained the sum of squares which is denoted here has adjusted of square.

So this is essentially in our terminology this is SS sum of a squares and this is in our terminology this is mean squares and this is value of F- statistics and this is the P-Value, now you can recall that we had divided the total sum of a squares in the case of analysis of

variance in two to components sum of square due to regression and sum of a squares due to error or sum of the square due to residual.

So this quantity here corresponding to this, this is giving us the value of SS res, that is the some of the square due to regression, and this is the quantity here which is trying Sto give us the value of SS res that is residual sum of squares. Now the software is also a dividing this some of the due to regression into two components which is some of the squares due to regression due to the variable x1, and this is the sum the square regression due to the variable x2.

And yeah, and here also you can see this is the sequential sum of the square which they again divided it into two components due to x1 due to x2 and they are using it in understanding the contribution of x1 and x2 in the regression sum of the square, so here you can see this value is 91.54% that means out of the two variable temperature and humidity, temperature is contributing nearly 91.54% and humidity is contributing only 5.19%.

Right and the component of error here is only 3.27%, right, the sum of squares due to residual will remain the same as SS res, right. So if you try to see even at these three values the make sense whenever the temperature is high usually the demand of water will increase, but if the humidity is also high the demand of water also increases, but the rate of change is low.

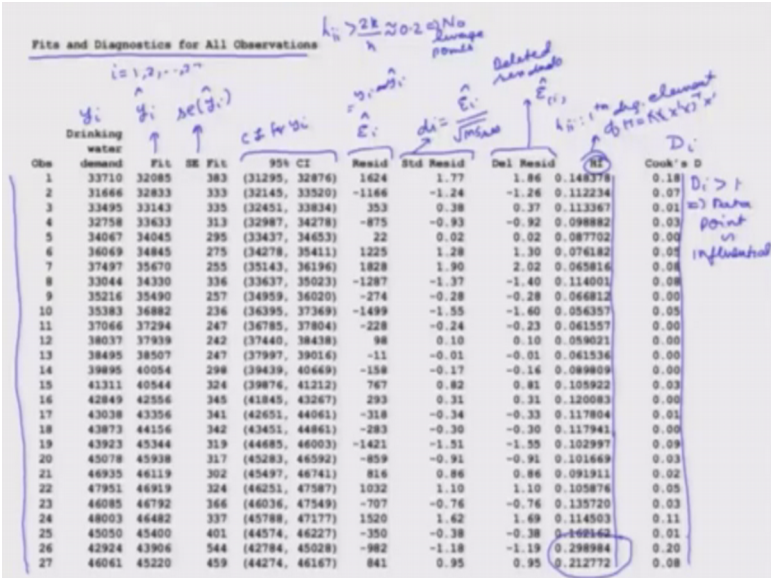
So that make sense that the model which we are fitted and the values which we have obtained here they are making some sense. Now this value here is mean square due to regression, and this value here is, this is mean square due to residuals okay, and the same thing here they have obtain the mean square due to regression due to x1 and here they have obtained it mean square due to regression due to the second variable x2.

And this value is the value of F-statistics that was obtain as MS res divide by MS res, so corresponding to this thing the p value is coming out to be 0.00, okay and here this the value of SST, this the total some of squares, and here you can see the degrees of freedom also here it is two, which was =k-1 that is 3-1, and this degrees of freedom was n-1, so this 27-and this was a n- k, so this is 24 degrees of freedom.

Now if you recall what was our objective in analysis of variance we were interested in testing the hypothesis like h1 beta1=beta2=0 and here we have chosen alpha = 0.05. So we can see here that the p value which we have obtain is closed to0 zero and this is smaller than alpha, so this implies that h naught is rejected.

So this means beta1 and beat2 are not equal to0 and correspondingly the variables x1 and x2 cannot be 0 simultaneously, so that is again confirming our outcome that we have got from the t values.

Refer Slide Time: (42:55)



Fits and Diagnostics for All Observations

Obs	Drinking water demand	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D
1	33710	32085	383	(31295, 32876)	1624	1.77	1.86	0.148378	0.18
2	31666	32833	333	(32145, 33520)	-1166	-1.24	-1.26	0.112234	0.07
3	33495	33143	335	(32451, 33834)	353	0.38	0.37	0.113367	0.01
4	32758	33633	313	(32987, 34278)	-875	-0.93	-0.92	0.098882	0.03
5	34067	34045	295	(33437, 34653)	22	0.02	0.02	0.087702	0.00
6	36069	34845	275	(34278, 35411)	1225	1.28	1.30	0.076182	0.05
7	37497	35670	255	(35143, 36196)	1828	1.90	2.02	0.065816	0.08
8	33044	34330	336	(33637, 35023)	-1287	-1.37	-1.40	0.114001	0.08
9	35216	35490	257	(34959, 36020)	-274	-0.28	-0.28	0.066812	0.00
10	35383	36882	236	(36395, 37369)	-1499	-1.55	-1.60	0.056357	0.05
11	37066	37294	247	(36785, 37804)	-228	-0.24	-0.23	0.061557	0.00
12	38037	37939	242	(37440, 38438)	98	0.10	0.10	0.059021	0.00
13	38495	38507	247	(37997, 39016)	-11	-0.01	-0.01	0.061536	0.00
14	39895	40054	298	(39439, 40669)	-159	-0.17	-0.16	0.089809	0.00
15	41311	40544	324	(39876, 41212)	767	0.82	0.81	0.105922	0.03
16	42849	42556	345	(41845, 43267)	293	0.31	0.31	0.120083	0.00
17	43038	43356	341	(42651, 44061)	-318	-0.34	-0.33	0.117804	0.01
18	43873	44156	342	(43451, 44861)	-283	-0.30	-0.30	0.117941	0.00
19	43923	45344	319	(44685, 46003)	-1421	-1.51	-1.55	0.102997	0.09
20	45078	45938	317	(45283, 46592)	-859	-0.91	-0.91	0.101669	0.03
21	46935	46119	302	(45497, 46741)	816	0.86	0.86	0.091911	0.02
22	47951	46919	324	(46251, 47587)	1032	1.10	1.10	0.105876	0.05
23	46085	46792	366	(46036, 47549)	-707	-0.76	-0.76	0.135720	0.03
24	48003	46482	337	(45788, 47177)	1520	1.62	1.69	0.114503	0.11
25	45050	45400	401	(44574, 46227)	-350	-0.38	-0.38	0.160162	0.01
26	42924	43906	544	(42784, 45028)	-982	-1.18	-1.19	0.298984	0.20
27	46061	45220	459	(44274, 46167)	841	0.95	0.95	0.212772	0.08

Okay, so now this is a table where we have stored all the fits and diagnostics and since we had given the option that we needed for all the observation so now we have here 27 observation and these statistics are available all the 27 observation, so now I will try to explain it in our symbolic notation what we have used earlier in the lecture, the first column is drinking water demand so this is nothing but our yi.

And this is here the fit that means this the value of yi hat, and this is the standard error of fit, so this is essentially standard error of yi hat, i goes from here 1 to 27 and this is here the confidence interval for yi and these are our ordinary residuals and which we had denote in our symbolic notations as epsilon i hat that we had obtain has yi minus yi hat or the difference between y i and y i hat.

And these are our standardized residuals that we had denoted in our symbolic language as di which was epsilon i hat divided by MS res, and this is our deleted residuals, and which we

had denoted epsilon i bracket hat, and this hi it is the in our notation is hii and this is the ith diagonal element of hat matrix h x x transpose x whole inverse x transpose, and this our Cook's D statistics and so this in our notation we also have denote it has here d or d's values are indicated by di.

Right, now we try to look at the interpretation of hi, so you may recall that we had concluded that if hii is greater than two k upon n which was a sort of guideline in our case this is 0.2 this would indicate that there are no leverage points. So you can see here in these values none of the values is very close 2.2 expect here these values over here.

Right, and then we will see later on also that these two values are quite far away from the fitted regression line so this is also well captured by our defined statistics. Okay, now we come to the interpretation of di's that is the Cook's D statistics. We had made a conclusion that that in case if di is greater than one then we would conclude that the data point is influential and it needs your attention.

But you can see here that none of the value here is close to one, so there is no question of having outlier and this was also indicated by our scatter diagram also that there are no such influential point which are trying to drag the fitted regression line towards them.

(**Refer Slide Time: 47:14)**

Obs	DFITS
1	0.775541
2	-0.447821
3	0.132147
4	-0.305984
5	0.006960
6	0.373087
7	0.536016
8	-0.502803
9	-0.074778
10	-0.391293
11	-0.059359
12	0.024961
13	-0.002987
14	-0.051316
15	0.278732
16	0.113801
17	-0.121985
18	-0.108689
19	-0.525999
20	-0.305484
21	0.272248
22	0.379276
23	-0.300406
24	0.606190
25	-0.166073
26	-0.776702
27	0.494432

$$DFITS = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{\hat{\sigma}^2_{(i)} h_{ii}}}$$

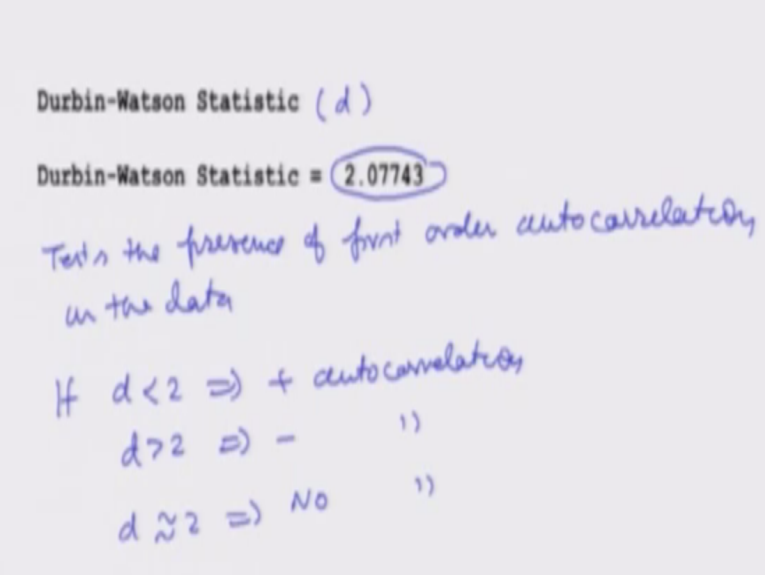
If $|DFITS| > 2\sqrt{\frac{k}{n}} \approx 0.66$ then the pt. warrants attention

Right, okay, now we come to our next data, here it is trying to give us the value of the DFITS statistics and you might call that we had defined the DFITS statistics as yi minus yi hat based on deleted residuals divided by sigma i square hat, based on deleted residuals hii and we had

conclude that if absolute value of DFITS is greater than twice of square root of k by n which in our case comes out to be 0.66 then ith point warrants attention.

So you can see here that here only there are two values here which are closed to or greater than point six, so here we have to be little bit careful that we will also see in the scatter diagram that two at the points at the end they are going towards downward direction, so this phenomenon is again correctly capture by our statistics and our regression analysis.

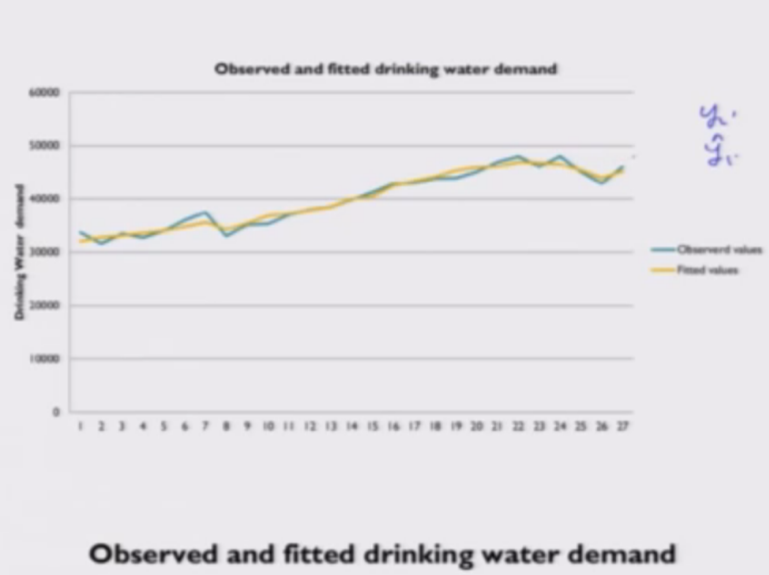
(**Refer Slide Time: 48:32)**



Right, now the next value is about Durbin-Watson statistics this is usually denoted by here d and Durbin-Watson statistics actually tests the presence of first order autocorrelation in the data and here the decision rule is very simple that if d is a smaller than two this indicates positive autocorrelation and if d is greater than two this indicates negative autocorrelation auto if d is close to two this indicates no auto correlation.

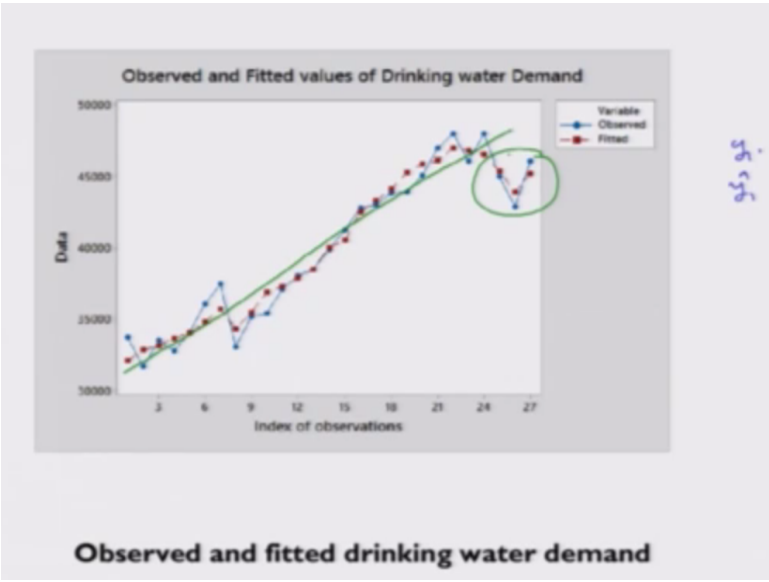
So now here in this case you can see that the value of Durbin-Watson statistics is close to two, so there is no autocorrelation of first order present in the data

(Refer Slide Time: 49:31)



Right, now we try to analysis these values through our graphical procedures so here I have plotted the values of yi and say yi hat, which I have obtained after fitting the model. One can see here that there is not much difference between the two values, so I can conclude that my model is well fitted and I even if I try to think it to use it for the prediction it is not difficult to say this model will continue further like this. Okay, and similarly in the next graphic also I have plotted these yi and yi hat point by point.

(Refer Slide Time: 50:07)



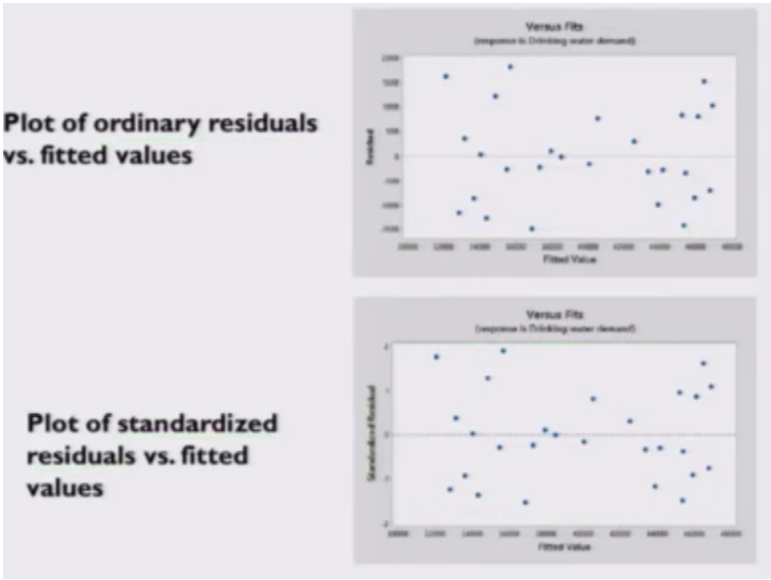
So you can see her say from here to here there is a sort of linear trend and towards the end part over here this is showing a sort of trend towards the downward and that was actually captured in the d statistics and say residuals. Okay so this is also confirming about the properties of the data, right. Now this is the normal probability plot.

(Refer Slide Time: 50:44)

And first figure is based on the normal probability plots based on ordinary residuals and second figure this the normal probability plot using the standardized residual. So you may recall that we had concluded that if all such points which are lying over here as a blue dots if they are lying closed to the line nearly on a straight line then we can be sure that the observation are coming from a normal population.

And this phenomena is happening in both the cases whether we are using the ordinary residual or the standardized residual so this gives us an assurance that the assumption what we have made that this correct that our observation are originating from normal distribution.

(**Refer Slide Time: 51:34)**



Now these are the plot of residuals versus the fitted value, so here in the first case I have taken the ordinary residual and in the second case I have taken the standardized residuals, and

you can see that there is not much difference between the two. Right and from these two pictures we can conclude that all the points possibly they can be contained in a sort of horizontal band and there no pattern like outward opening funnel or inward opening funnel or there are no indication of having non linear relationship.

So from these two graphics we can conclude there is no problem of heteroscedasticity in the data and our variance remains constant. Okay now I have consider here an example and I have explained the complete statistical analysis whatever we have done in the earlier lectures in all details. Now please try to have a look at all these results, and try to understand yourself that which of the quantity is indicating what and how it is computed, once you know this thing then you will have more inside into data analysis.

Particularly in those situation when there is a problem because you are not going to deal in practice with 27 observation, but there can be several hundred observations and then these diagnostics these statistical results will help you in taking a decision that whether the fitted regression model is good or bad, you may not have an opportunity to look at the individual data points, and you would like to inspect the individual data points only in case of any problem.

So I stop here and in the next lecture I will try take one more example to illustrate some further aspects of multiple linear regression analysis, till then good bye.