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

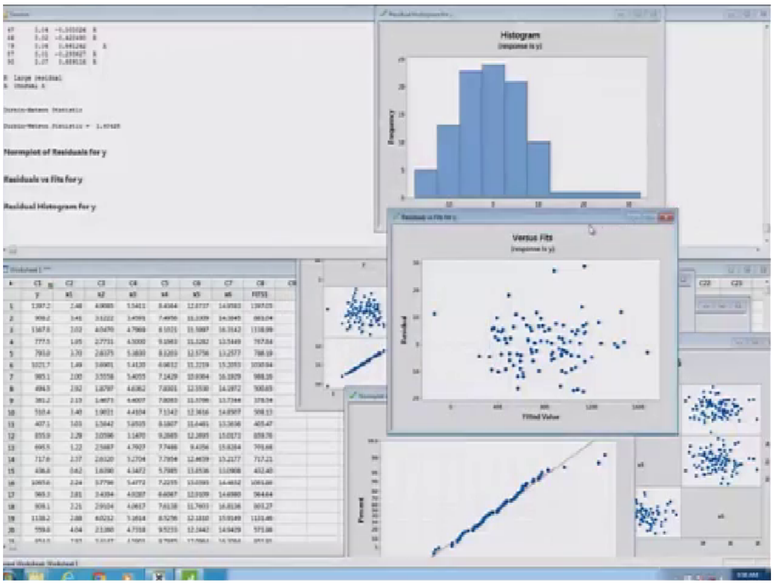
# Lecture-20
# Software Implementation of Multiple Linear Regression Model using MINITAB (continued)

Welcome to the lecture you may recall that in the earlier lecture we had considered an example in which we had considered some dataset and we had expose it to a statistical software and after that we had analysis the statistical outcome. We had learnt how to interpret the different components of statistical outcome and based on that how are going to conclude about the statistical model in that example you may recall that everything was nice.

Means as soon as get the model everything was coming out exactly in the same way as we had studied in the lecture. In practice this is always not possible. A good statistical model is not usually obtainable in a first attempt rather it is a sort of iterative process. We start with the data, we try to investigate it from different perspectives and then we try to get a model by looking at the outcome of the statistical software we try to decide whether the model, which we are obtained is good or bad.

Does this satisfy all the assumptions of the linear regression model, is there any problem in the outcome and based on that we try diagnose all the problems in the model, then in the second step we try to correct it and then were conduct the statistical analysis one again and this process continues till we obtain a good statistical model. So in order to illustrate this type of aspects so I have considered a dataset of size hundred.

**(Refer Slide Time: 06:46)**



```
R  Large residual

Durbin-Watson Statistic

Normplot of Residuals for y

Residuals vs Fits for y

Residual Histogram for y
```

Size hundred, I am taking because that's till bit large dataset earlier we are taken very small data set in which you can see everything from the eyes very easily. So now we are going to consider some other complexities which are involved in the dataset. So now let as try to come on this example and if you try see here I already have entered this data in the Minitab software, so I am considering here six independent variables and its outcome is given here in the column as y.

So now first of all we have to investigate that whether a linear regression model can be fitted over here or not so you may see that in the earlier case we had only three variables so I can make different types of thing, but here now we have six variables so things are not so straight forward, but still we make an attempt. We try to make here a matrix scatter plot, mean I will try to choose here all the variables.

And if you try to see this matrix scatter plot comes out be quite clumsy, but still you can see here that it is trying to give us some information, but it is pity difficult here, so what I try to do here I tried to make or consider two variables at a time. So I try to make a scatter plot once again and I try to considered only x one and x two, and one can obtain scatter plot like this.

Similarly, I tried to considered two variables at that time, now I tried to considered y and x three and x four and we get here a plot like this one and similarly I tried to make here another graph with the variables x5 and x6 y versus x5 and x6 and I tried to bring these graph in a separate sheets so that we can discuss it later on. After this we tried to fit here a model and we tried to give the continuous predicator here has x1, x2, x3, x4, x5, x6.

Now we have to give here different types of options so for example here I am including the intercept term in the model and we are going for different types options that here I am going to consider the level of significance to be 5% so the confidence level for all the interval is at ninety 5% over here, and we all going to considered a two-sided test of hypotheses and we are going to consider the adjusted or simple sum of squares.

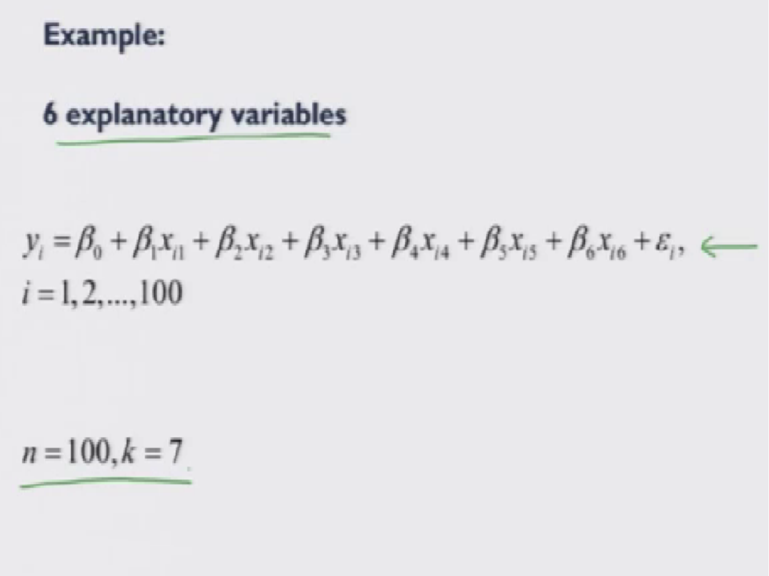
Now we would like to have different types graph we can have a Histogram of residuals, Normal probability plus Residuals versus fit and here would try to see that I am also considering here the Histogram of residuals, this graphic will make here sense becomes we have hundred observation in the earlier case we had just observations so it does not make much sense or it does not give good information to us.

And for the result I would try have all the things, method, analyses of variance, model summary, coefficient. regression equation, fits and diagnostic and this I am now going only for as usual observation, because there are hundred observation otherwise we will get the very long table so we are interested only in the unusual observation and we would like to have the Durbin-Watson test statistics.

Now lets us try to obtain the result, so you can see here that this is histogram that we are obtaining here and this is a sort of residual versus fit that is obtain over here and similarly here we obtain another graphics like as here, this is the residual versus fitted value and Similarly we have obtained here the normal probability plot. Now if you try to look at the outcome, so we have obtain here the analysis of variance table and here we obtain the model summary then we have obtain here the coefficients.

Finally we have the regression equation here and similarly these are some statistics for the unusual observations and then finally we have Durbin–Watson test statistic, so I just copy all the things in a separate sheet and then we try to understand the entire analysis?

**(Refer Slide Time: 07:49)**



**Example:**

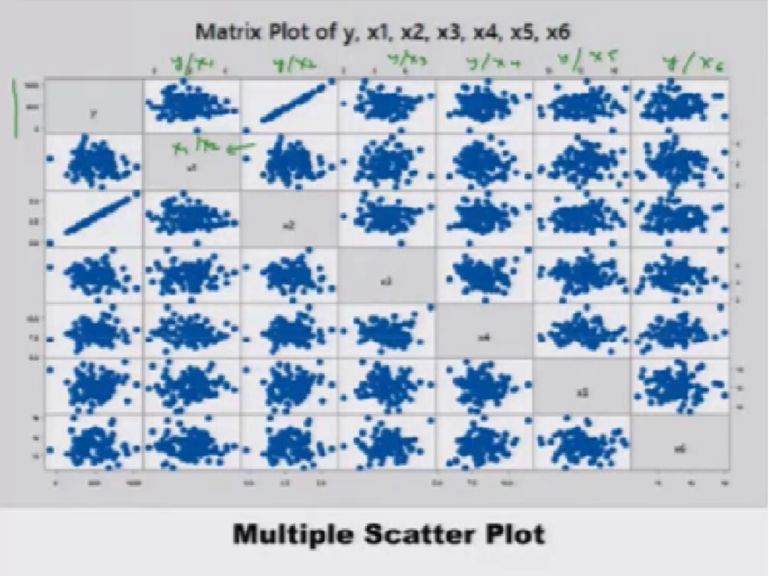
**6 explanatory variables**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \varepsilon_i, \longleftarrow$$
$$i = 1, 2, \ldots, 100$$

$$n = 100, k = 7$$

So now you can see here that we are considering here linear regression model with six explanatory variables, which are denoted by a x1, x2, x3, x4, x5 and x6 and based on that we have written this model and we are considering here hundred observation so n=100 and the number of independent variables including the intercept term that is here now k it is a seven.

**(Refer Slide Time: 08:23)**



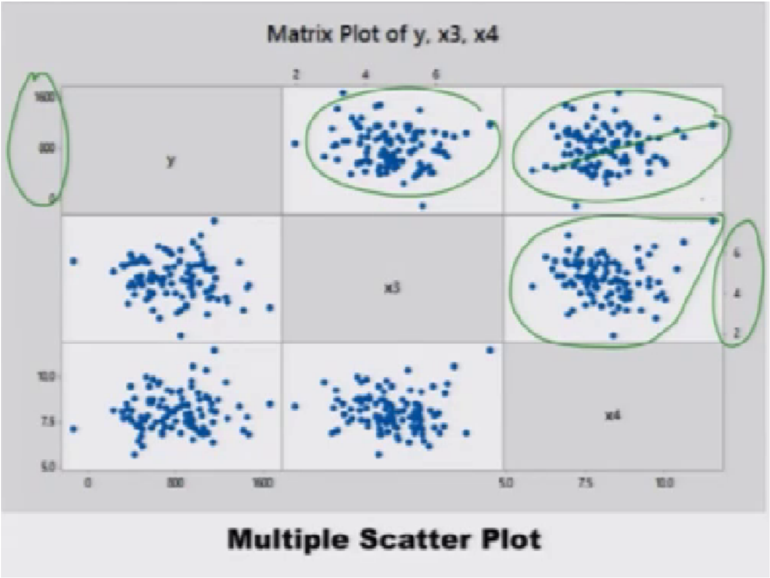
Matrix Plot of y, x1, x2, x3, x4, x5, x6

**Multiple Scatter Plot**

So now this is the matrix is scatter plot that we had obtain earlier you can that here it is quite clumsy and it is difficult to get a very clear information all though you can see here that this is giving a plot between y and x1, this is between y and x2, this is between y and x3, this is between y and x4 this is between y and x5 and this is between y and x6 and these are the plots, for example this is the plot between x1 and x2 and so on.

And we also have to take care of the scale also, so that we can find out whether there is really a linear pattern or not, so now you can see the first complicity when you are dealing with higher number of independent variables and large number of dataset. So we try to use here one thing that we try to consider this matrix plot with two variables at a time.

**(Refer Slide Time: 09:20)**



Multiple Scatter Plot

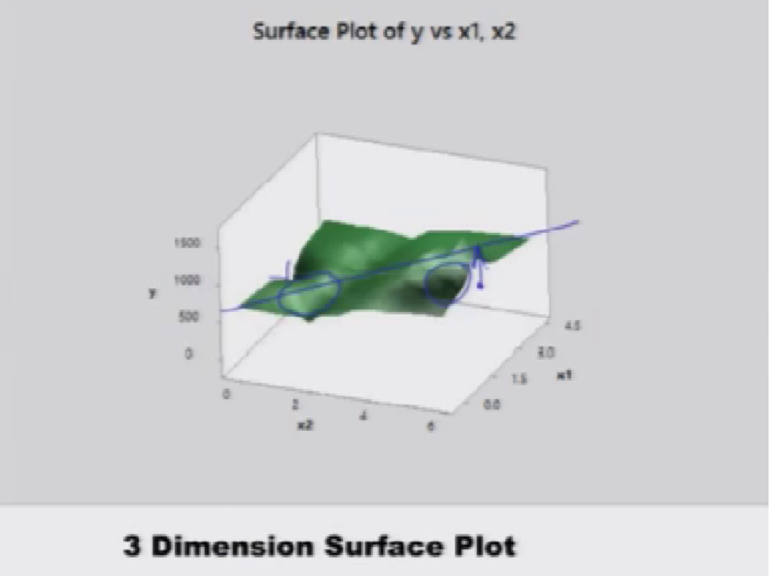
So I try to plot here y with respect to x1 and x2 and we plot this diagram and yes, that looks quite informative has compared to earlier, but you can see here the range here is between 0 to 1600. So that is while this is not giving us a very clear inference but still we can see here that is not actually 100% random there seems to be some trend over here.

When we trying to look at this one this is clearly giving us a linear trend, so there is no issue and when we are trying to look at the scatter diagram between x1 and x2, here the range is not actually much between 0 and 4 and that is clearly indicating that x1 x2 are independent, similarly when we try to look at the next matrix plots which is between y x3 and x4, we have a similar thing again the range here is very, very high in this case.

And in this case also the range is very, very high this is here is 0 to 1600, but still we can see that there seem to be some approximate linear trend over here and an important part is that here the range is not that high it is between only 2 and 6 and this scatter diagram is indicating that x3 and x4 are also independent and similarly if you try to considered the matrix is scatter plot of x5 and x6 there is again similar to other plots.

So here you can see although we are considering two variables at time but still that is not giving us a very clear picture that what is really happening we are not 100% convinced. So what we try to do here that we had plotted the surface plots. The surface plots we had considered in the earlier lecture that they can be obtain in the software Minitab just by click and in other software also they can be obtain very easily.

**(Refer Slide Time: 11:36)**



One can see here that there is a linear trend, so one can see that and the different shades here for example here can see the shade is quite dark whereas you can see here the shade here quite light, so they are going to give us the folds of the surfaces. For example here you can see that there is a dip and here also there is a sort of up. But still more or less it is showing that yes, it means that relationship of y with the respect to x one and x two jointly is not that bad.

Similarly when I try to plot here y with respective x3 and x4, so you can see here there are some sharp peaks over here just like a small mountains, and this type of perturbation they are trying to give us an idea about the disturbances on the surface, but still one can see here that there is a sort of linear trend over here. So we can be confident that there is a sort a sort joint relationship x1 and x2 with respective y is not bad and it is approximately linear.

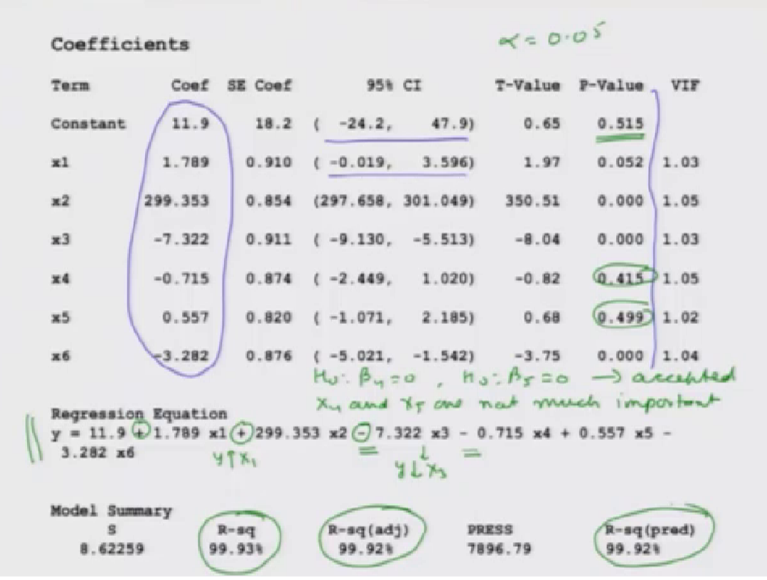
And similarly when we try to plot y with the respective x5 and x6 we can see that here this perturbations are little bit higher and the peaks and the depth of the surface is like as here in the peak and here the depth and the depth for example here the peak they are pretty high and they are higher than such deviations in comparison to x1, x2, and x3 and x4 and here also we can see here there is a sort of increasing linear trend.

But still you have to keep in mind that these are the relationship of two independent variable with respect to response variable and there are three possible outcomes and similarly we have to make such figures with the respect to all possible combination like as x1, x2, x4 x5, and then based on that we have take a finally decision by combining all the outcomes together that whether the relationship is going to linear are not.

So I had plotted all the things I am skipping all those figures, but I have taken these three figures to illustrate the methodology, but finally I can conclude, yes, I can fit here a linear regression model, and also remember one thing these are the individual relationship of two variables with the response variable at a time, but we are actually interested in the joint relationship of six explanatory variables with respect to y.

So when we find that pair-wise relationship of independent variables with the respect to the response variables are quiet linear then we can expect that there linear combination will also be linear.

**(Refer Slide Time: 15:00)**



$\alpha = 0{\cdot}05$

```
Coefficients

Term        Coef  SE Coef         95% CI          T-Value  P-Value   VIF
Constant    11.9     18.2  (  -24.2,     47.9)       0.65    0.515
x1         1.789    0.910  ( -0.019,    3.596)       1.97    0.052  1.03
x2       299.353    0.854  (297.658,  301.049)     350.51    0.000  1.05
x3        -7.322    0.911  ( -9.130,   -5.513)      -8.04    0.000  1.03
x4        -0.715    0.874  ( -2.449,    1.020)      -0.82    0.415  1.05
x5         0.557    0.820  ( -1.071,    2.185)       0.68    0.499  1.02
x6        -3.282    0.876  ( -5.021,   -1.542)      -3.75    0.000  1.04

Regression Equation
y = 11.9 + 1.789 x1 + 299.353 x2 - 7.322 x3 - 0.715 x4 + 0.557 x5 -
 3.282 x6

Model Summary
      S     R-sq  R-sq(adj)    PRESS  R-sq(pred)
8.62259   99.93%     99.92%  7896.79      99.92%
```

$H_0: \beta_4 = 0$ , $H_0: \beta_5 = 0$ → accepted

$x_4$ and $x_5$ are not much important

$y \uparrow x_1$

$y \downarrow x_3$

So this is how we try to taken a final judgement, because final judgement as to be in terms of yes or no whether I can fit a linear regression model or not. Now first I try to consider the outcome of software in the form of regression coefficient. I am not going to discuss here now all the components because we had discussed it in the earlier lecture and this is the similar software outcome.

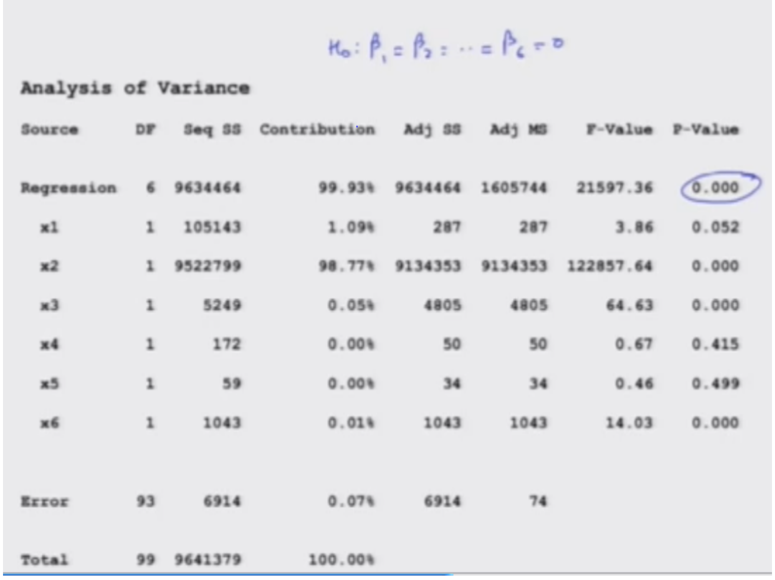
So these are here the regression coefficients that are obtain and these are here, their respective confidence interval, and so first we try to look at here, look at the P values, and we can see here that we had consider the alpha to be 0 point 0 5. So now there are three values one is here intercept term and two values here corresponding to x4, and x5 which are higher than . 05.

So we can conclude that when we are trying to test the hypotheses h naught beta4=0 and h naught beta5=0, both are going to be accepted. So we can conclude here now that x4 and x5 are not much important. In the sense that they are not contributing significantly in explanting the variation in Y, and we are obtain this model over here, one thing you have to notice here, that there are some terms which have got a positive sign, and there are some terms which have got a negative sign.

So we have understood that the meaning of the plus sign is that the relationship between y and x1 is increasing, so y is increasing as x one increases, but similarly for the x3 here the sign here is negative, so that mean the relationship y and x3 is decreasing, so as x3 increases the outcome of y decreases. Now when we look at the model summary we can see that the R square is coming out pretty high, and even the R square prediction is also pretty high.

So the fitted model is good, but there is a doubt that x4 and x5 variables are not contributing much and similarly is the doubt for the presence of intercept term also here. Right, so we try to now consider the analysis of variance part, so here we are essentially interested in testing the hypotheses h naught beta1=beta2 up to here beta6=0.

**(Refer Slide Time: 17:47)**

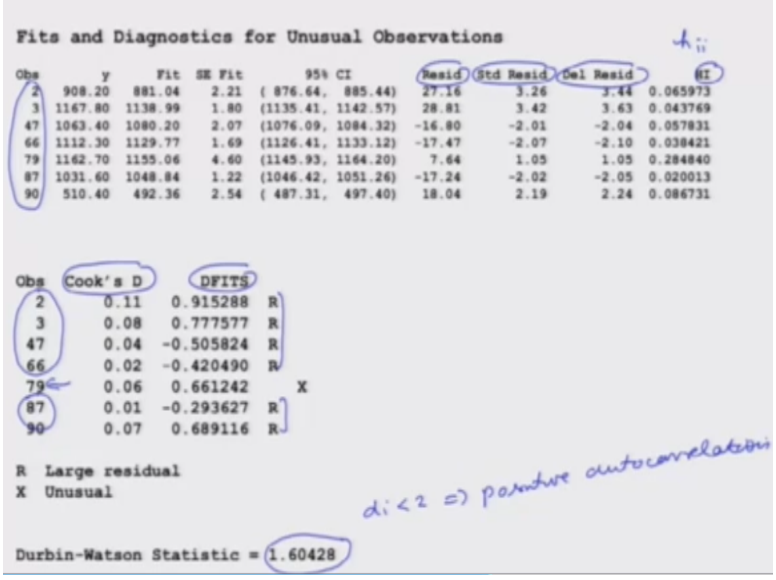


$H_0: \beta_1 = \beta_2 = \cdots = \beta_6 = 0$

```
Analysis of Variance

Source       DF   Seq SS  Contribution   Adj SS   Adj MS   F-Value  P-Value

Regression    6  9634464        99.93%  9634464  1605744  21597.36    0.000
  x1          1   105143         1.09%      287      287      3.86    0.052
  x2          1  9522799        98.77%  9134353  9134353 122857.64    0.000
  x3          1     5249         0.05%     4805     4805     64.63    0.000
  x4          1      172         0.00%       50       50      0.67    0.415
  x5          1       59         0.00%       34       34      0.46    0.499
  x6          1     1043         0.01%     1043     1043     14.03    0.000

Error        93     6914         0.07%     6914       74

Total        99  9641379       100.00%
```

So you can see here that the corresponding P value is coming out to be close to0, so one can conclude that at 5% level of significance that none of the individual variable is close to 0, but again you can see now this is contracting between the outcome that we had obtain early that x4 and x5 are not contributing much, so now we try to understand that how to conclude finally.

**(Refer Slide Time: 18:39)**



```
Fits and Diagnostics for Unusual Observations

Obs        y      Fit  SE Fit          95% CI          Resid  Std Resid  Del Resid        HI
  2   908.20   881.04    2.21  ( 876.64,   885.44)    27.16       3.26       3.44  0.065973
  3  1167.80  1138.99    1.80  (1135.41,  1142.57)    28.81       3.42       3.63  0.043769
 47  1063.40  1080.20    2.07  (1076.09,  1084.32)   -16.80      -2.01      -2.04  0.057831
 66  1112.30  1129.77    1.69  (1126.41,  1133.12)   -17.47      -2.07      -2.10  0.038421
 79  1162.70  1155.06    4.60  (1145.93,  1164.20)     7.64       1.05       1.05  0.284840
 87  1031.60  1048.84    1.22  (1046.42,  1051.26)   -17.24      -2.02      -2.05  0.020013
 90   510.40   492.36    2.54  ( 487.31,   497.40)    18.04       2.19       2.24  0.086731


Obs  Cook's D      DFITS
  2      0.11   0.915288  R
  3      0.08   0.777577  R
 47      0.04  -0.505824  R
 66      0.02  -0.420490  R
 79      0.06   0.661242     X
 87      0.01  -0.293627  R
 90      0.07   0.689116  R

R  Large residual
X  Unusual

Durbin-Watson Statistic = 1.60428
```

$h_{ii}$

$d_i < 2 \Rightarrow$ positive autocorrelation

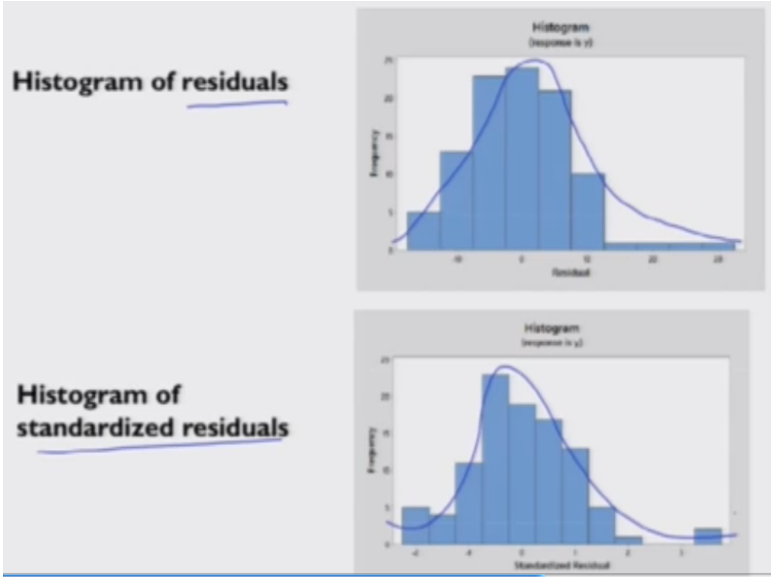
So next we consider different types of fits and diagnostic they have been obtained only for the unusually observations so you can see here that the software is diagnosing that this many observations they are looking to be unusual and they have obtained different types of residual standardized residual deleted residuals and this is value of here hii with ith diagonal value of the hat matrix h and Cook's distance over here and DFFITS statistics.

So, this we already have understood in the earlier lecture that how to interpret these things, but based on that now we have an idea, that observations number here 2, 3, 47 and 66 and 87 and 90 they are the large residual, which are denoted by here R and observation number here seventy nine this is sum unusual observation, so now this is our responsibility to back into the data, and we look into the behavior of this observation.

And then we have to a final call weather they are really influential observation outliers are they are unusual observation or say leverage point. And one can also see here that Durbin-Watson statistics is coming out to be here 1.60, and this indicates that the rule of Durbin-Watson statistics that we have discussed earlier that if d is smaller than two then this indicates the presents of positive autocorrelation.

So one thing we have diagnose here that there is a possibility of presence of first order positive autocorrelation and this is possibly making this types of strange observations in the earlier analyses. So now what we do here that since we have now obtained that, that the two variables x4 and x5 are not good.

**(Refer Slide Time: 20:57)**



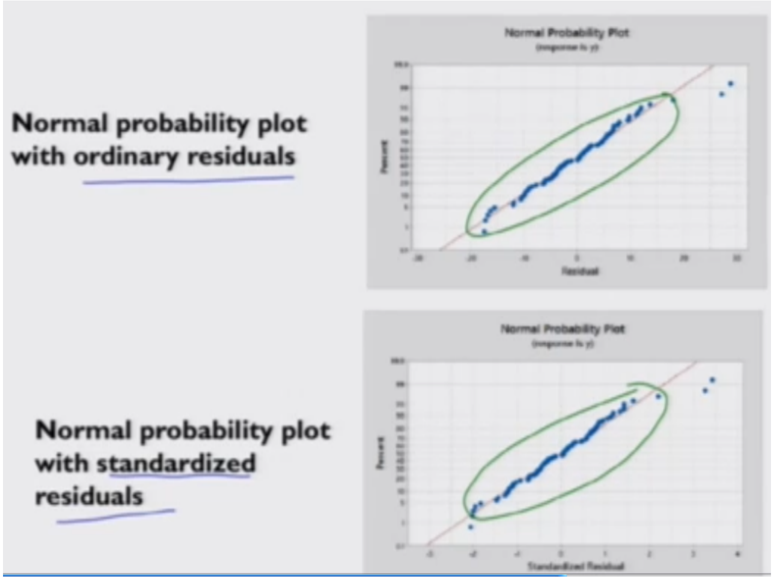
Histogram of residuals

Histogram of standardized residuals

So our next objective will be to conduct the analyses once again, by removing these variables but before that let us try to look at the different graphical outcomes of the same analysis. So here I have plotted the histogram of ordinary residuals and standardized residual so this residuals also have a normal distribution but since you can see here that since we have only here hundred observation.

So one fit here a curve like this one, but still that is not100% normal, but that we known regarding in case if we try to have more observation possibly this picture may tends towards normal curvature and similarly here also the picture looks like, so this it is not hundred percent normal, but it is giving a good confidence that if we have some more number of observation possibly this will picture will look more normal.

**(Refer Slide Time: 21:58)**

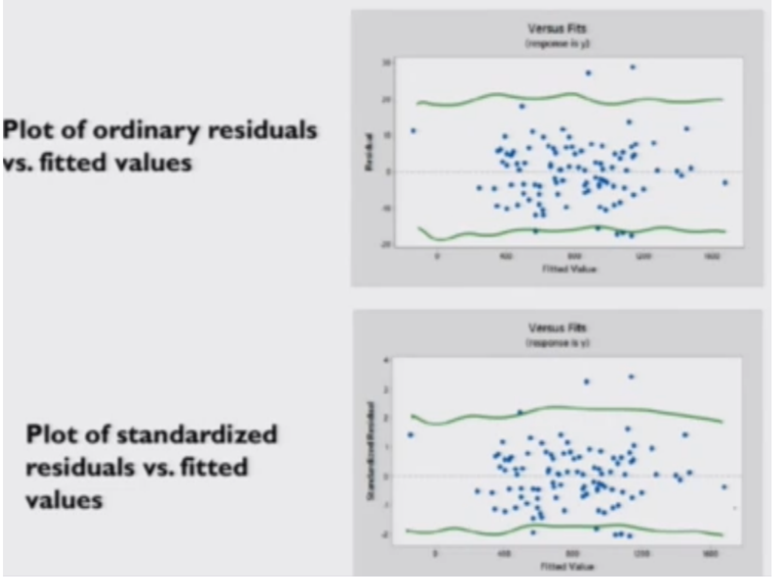


Normal probability plot with ordinary residuals

Normal probability plot with standardized residuals

More stronger conclusion can be obtain from the normal probability plot and this normal probability is plot is plotted here for the ordinary residuals and for the standardized residuals, and one can see here that all the points are lying here mostly around the straight line right, so one can conclude here that the observations are coming from a normal population.

**(Refer Slide Time: 22:29)**

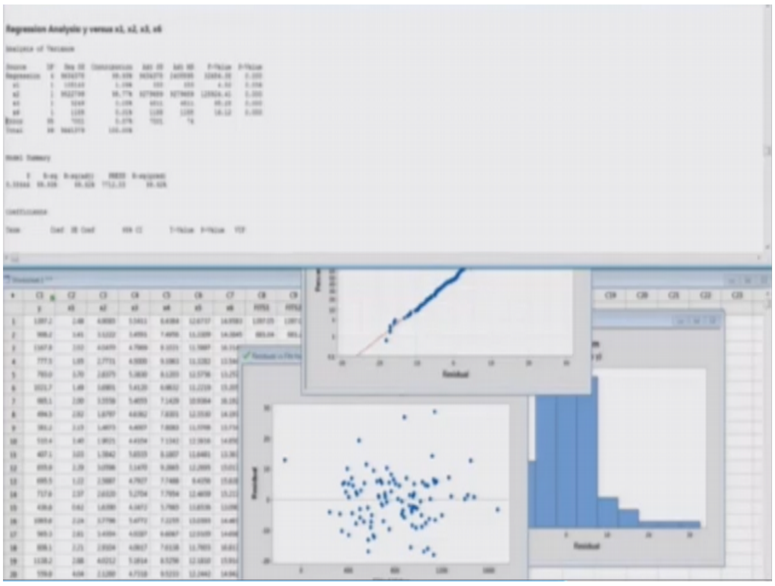


Plot of ordinary residuals vs. fitted values

Plot of standardized residuals vs. fitted values

Next we have made the plots of residual versus fit and their standardized residuals versus fit and one can see here that most of the observation they are enclosed in a sort of this horizontal band, right. So one can say that here that there is not much issue and the variance remains almost the constant, right. So now we have seen here that two independent variable x4 and x5 they are turning out to be not contributing towards the model.

**(Refer Slide Time: 23:19)**

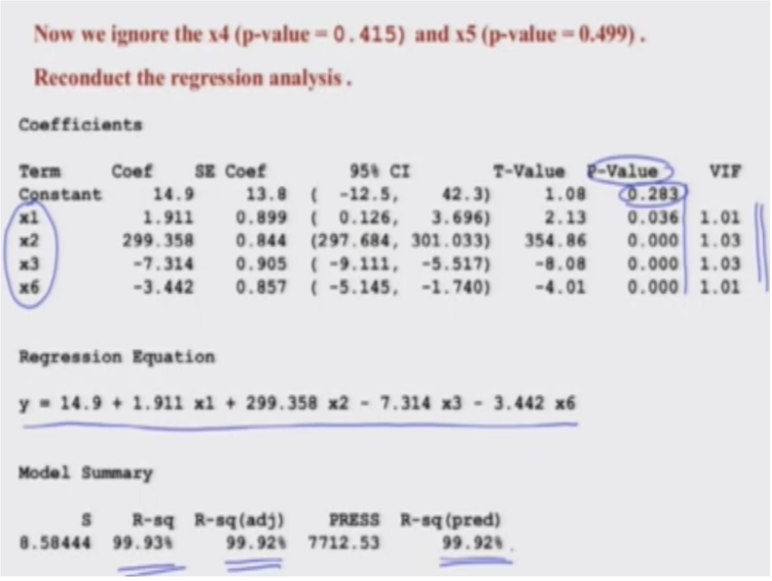


Regression Analysis: y versus x1, x2, x3, x6

So we try to conduct this regression analysis once again by ignoring the two variables, so we try to do it here again and we consider R linear regression of Y with respect to four variables only x1, x2 x3 and x6, right and we try to obtain here the similar thing you can see here we have obtain here the histogram of residuals, the plot of residuals versus fit and the normal probability plot over here.

And whatever are the observations that we have obtain here, the statistical out come this I am trying to copy and bring it to another sheets so that we can discuss about them.

**(Refer Slide Time: 24:05)**



Now we ignore the x4 (p-value = 0.415) and x5 (p-value = 0.499).

Reconduct the regression analysis.

```
Coefficients

Term        Coef   SE Coef          95% CI           T-Value  P-Value   VIF
Constant    14.9      13.8  (  -12.5,     42.3)        1.08    0.283
x1         1.911     0.899  (  0.126,    3.696)        2.13    0.036   1.01
x2       299.358     0.844  (297.684, 301.033)       354.86    0.000   1.03
x3        -7.314     0.905  ( -9.111,   -5.517)       -8.08    0.000   1.03
x6        -3.442     0.857  ( -5.145,   -1.740)       -4.01    0.000   1.01


Regression Equation

y = 14.9 + 1.911 x1 + 299.358 x2 - 7.314 x3 - 3.442 x6


Model Summary

      S    R-sq  R-sq(adj)    PRESS  R-sq(pred)
8.58444  99.93%     99.92%  7712.53      99.92%
```

Okay, so you can see here this is same outcome that we had obtain earlier, so now you can see that we are only consisting here x one, x2, x3 and x6, and now if you see here now the resulting p values here they are given in this column, and you can see that they are somehow little bit changed than earlier, but still this constant term is still indicating that the presence of consent term is possibly not desirable.

So that we have to explore further, but again that depends on the situation and the experimental conditions, if you remember the interpretation of the intercept term was that it is the value of average value of y when all the independent variables takes value 0.So the experimenter has to take final call that weather there has to been in intercept term or not, but any way we have re-conducted the analyses and finally we have obtained this model.

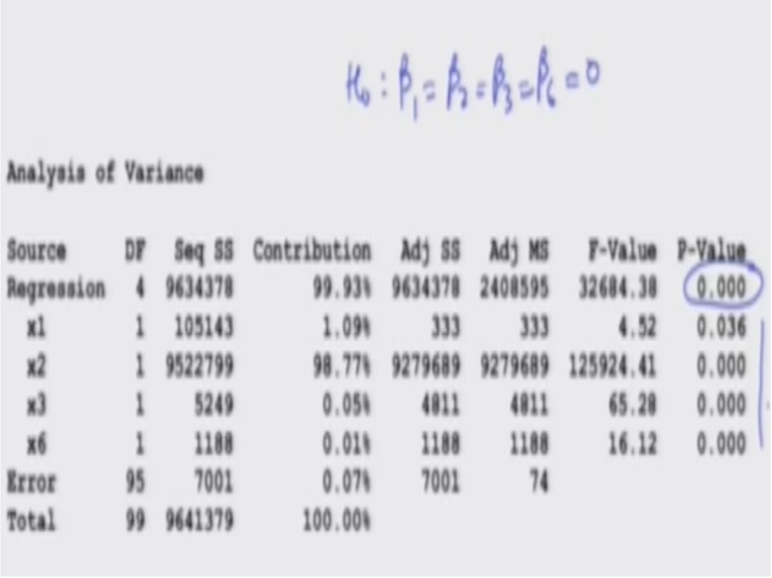
And you can see that this model and the values of regression coefficients they are pretty different than what we had obtained earlier, and this various inflation factor also they are very close to here one, so that is again indicating that there is no problem of multicollinearity in the sense that all the variables x1, x2 and x3 and x6, they are independent and the model summary is also indicating that the value of R square, R square adjusted they are pretty close to 99%.

And R square is prediction is also close to ninety nine percent because the model is really good, but before I go further, let me try to point out one thing, usually in experimental science many scientist and researcher they believe the outcome to be good only when R is square is pretty high say more than night five percent, but if you are really considering a real life data the possibilities of getting such a high values are not really very high.

Because the interpretation goes like this that when I say that the model is 99% good that mean it is really close to the true experimental set up, well if you get it that is very nice, but in real data survey it is difficult to get such a high value, So that is the choice of the experimenter to decide whether he considers the R square equal to point seven to be higher value or a lower value or even R square is equal point nine to be higher value or lower value.

So when we are doing a regression analysis it requires sum theoretical knowledge understanding of basic fundamentals, understanding experimental conditions and the condition under which the data has been collected, the role of those variables in experimental condition environment in which the data has been collected and finally this is the only experimenter who takes the final call, okay. So next we have obtained here the analyses of variance table.

**(Refer Slide Time: 27:37)**

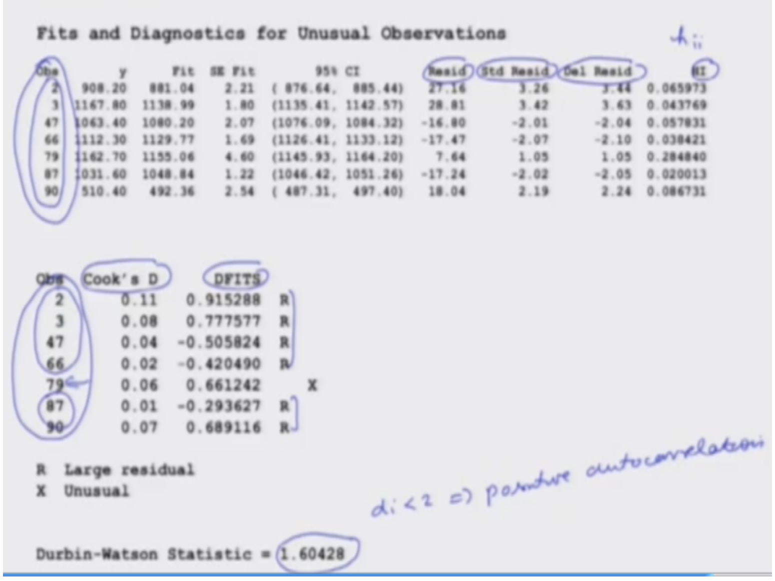


$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_6 = 0$

```
Analysis of Variance

Source      DF   Seq SS  Contribution   Adj SS   Adj MS    F-Value  P-Value
Regression   4  9634378        99.93%  9634378  2408595   32684.38    0.000
  x1         1   105143         1.09%      333      333       4.52    0.036
  x2         1  9522799        98.77%  9279689  9279689  125924.41    0.000
  x3         1     5249         0.05%     4811     4811      65.28    0.000
  x6         1     1188         0.01%     1188     1188      16.12    0.000
Error       95     7001         0.07%     7001       74
Total       99  9641379       100.00%
```

And so now in this case our h naught is going to be h naught beta1=beta2=beta3=beta6=0, and one can see here that this P value is coming out to be close to 0, all other value is are also because it is smaller the .05, so one can see here that the h naught is rejected and all the variable x1, x2, x3 and x6 they are contributing to us the model, in explaining the variation in the value of y.

**(Refer Slide Time: 28:15)**



```
Fits and Diagnostics for Unusual Observations

Obs        y      Fit  SE Fit          95% CI          Resid  Std Resid  Del Resid        HI
  2   908.20   881.04    2.21  ( 876.64,  885.44)      27.16       3.26       3.44  0.065973
  3  1167.80  1138.99    1.80  (1135.41, 1142.57)      28.81       3.42       3.63  0.043769
 47  1063.40  1080.20    2.07  (1076.09, 1084.32)     -16.80      -2.01      -2.04  0.057831
 66  1112.30  1129.77    1.69  (1126.41, 1133.12)     -17.47      -2.07      -2.10  0.038421
 79  1162.70  1155.06    4.60  (1145.93, 1164.20)       7.64       1.05       1.05  0.284840
 87  1031.60  1048.84    1.22  (1046.42, 1051.26)     -17.24      -2.02      -2.05  0.020013
 90   510.40   492.36    2.54  ( 487.31,  497.40)      18.04       2.19       2.24  0.086731


Obs  Cook's D      DFITS
  2      0.11   0.915288  R
  3      0.08   0.777577  R
 47      0.04  -0.505824  R
 66      0.02  -0.420490  R
 79      0.06   0.661242     X
 87      0.01  -0.293627  R
 90      0.07   0.689116  R

R  Large residual
X  Unusual


Durbin-Watson Statistic = 1.60428
```

$h_{ii}$

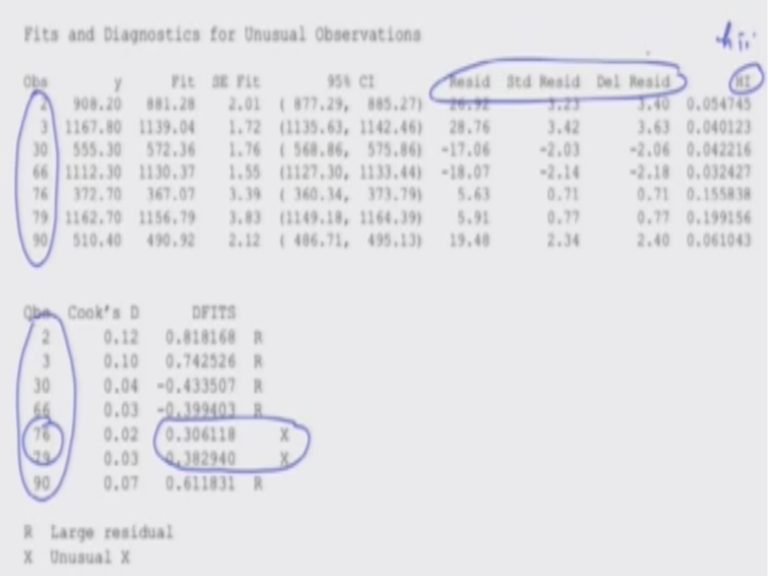
$d_i < 2 \Rightarrow$ positive autocorrelation

And here we have obtain the fits and diagnostics for the unusual observation and we seen earlier also that this many observation have been diagnosed to be either having a large residual are an unusual observation, but here you can see there is a change that now here there are two observations, which are indicating that they are unusual earlier they was only one observation which is unusual.

Well, these things are going to happen and that is why we need to revise our statistical analyses again and again till we reach to a good model. So I am not going into the details of this outcome, but the suggestion would be that we try to delete these observations from the dataset and we try to conduct the entire analyses once again. In fact a good suggestion is that after the first regression analyses for example we have obtain that these many observations have some trouble.

So a better option is to just conduct the entire analyses once again without deleting x4 and x5 and then we try to see the outcome and accordingly we have to take a suitable decision.

**(Refer Slide Time: 29.40)**

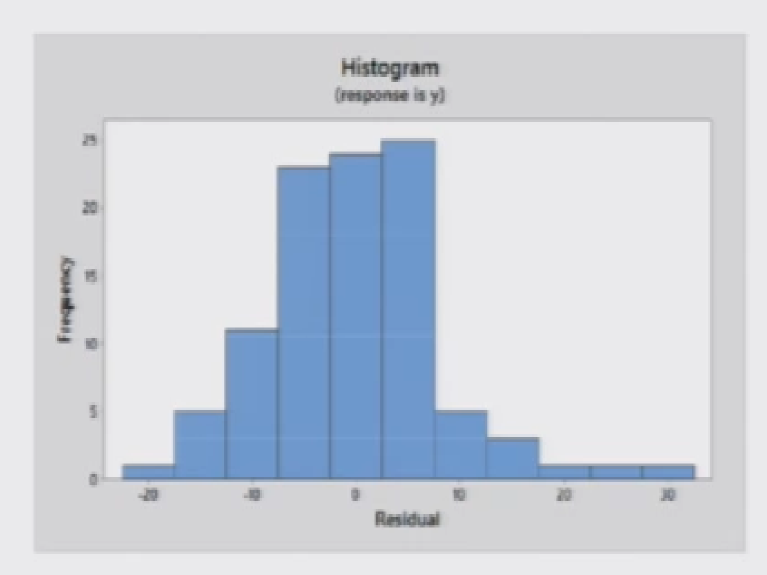


$h_{ii}$

```
Fits and Diagnostics for Unusual Observations

Obs        y      Fit  SE Fit          95% CI           Resid  Std Resid  Del Resid       HI
  2   908.20   881.28    2.01  ( 877.29,  885.27)       26.92       3.23       3.40  0.054745
  3  1167.80  1139.04    1.72  (1135.63, 1142.46)       28.76       3.42       3.63  0.040123
 30   555.30   572.36    1.76  ( 568.86,  575.86)      -17.06      -2.03      -2.06  0.042216
 66  1112.30  1130.37    1.55  (1127.30, 1133.44)      -18.07      -2.14      -2.18  0.032427
 76   372.70   367.07    3.39  ( 360.34,  373.79)        5.63       0.71       0.71  0.155838
 79  1162.70  1156.79    3.83  (1149.18, 1164.39)        5.91       0.77       0.77  0.199156
 90   510.40   490.92    2.12  ( 486.71,  495.13)       19.48       2.34       2.40  0.061043

Obs  Cook's D      DFITS
  2      0.12   0.818168  R
  3      0.10   0.742526  R
 30      0.04  -0.433507  R
 66      0.03  -0.399403  R
 76      0.02   0.306118      X
 79      0.03   0.382940      X
 90      0.07   0.611831  R

R  Large residual
X  Unusual X
```

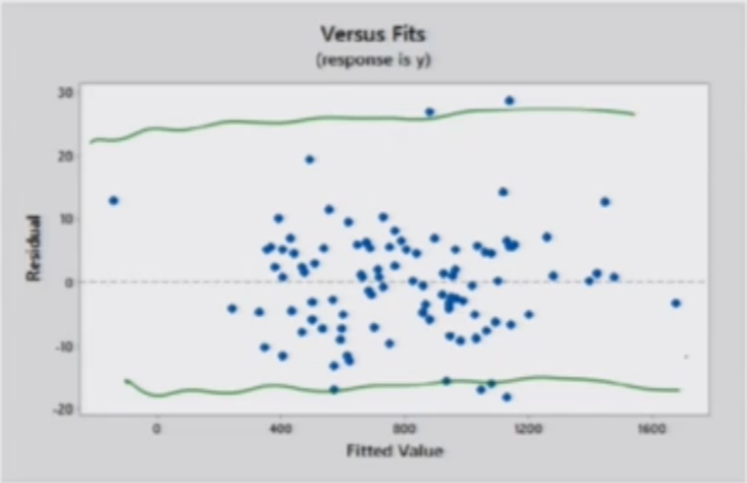
Well, now here based on the value of HII deleted residual, standardized residual, residuals you can also take a conclusion and you can decide that which of the observations are going to be influential outlier or anything else.

**(Refer Slide Time: 29:59)**



Right so now we have plotted different types graphical diagnostic and here also you can see as earlier there is no issue of here normal t assumption but here are the sake of understanding I have plotted only here one histogram of only the standardized residual and similarly we have obtain diagnostic or a graphically presentation of residuals versus fits, but we have consider only the ordinary residuals.

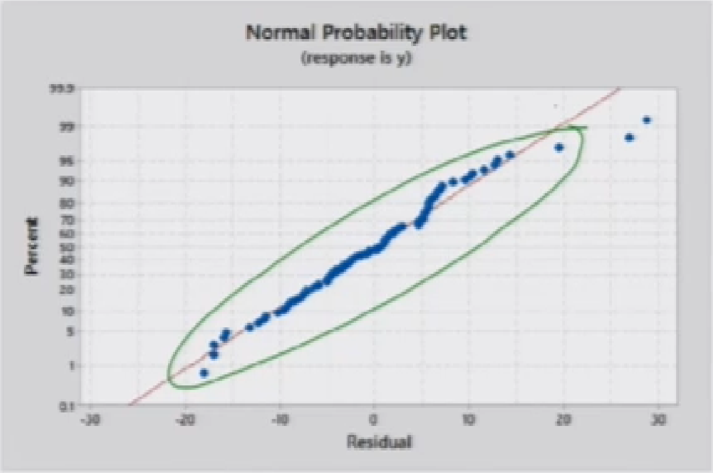
**(Refer Slide Time: 30:20)**



**Plot of ordinary residuals vs. fitted values**

And we can see here that all this points are nearly in close inside the horizontal band, so the assumption like constant variance and other thing they are satisfied here.

**(Refer Slide Time: 30:45)**



**Normal probability plot with ordinary residuals**

And we have obtain the normal probability here for the ordinary residuals and we have obtain that this that most of the points a lying near the straight line, so this gives as a sort of assurance that the observations are coming from a normal population. So now from this example you can see that whenever we trying to do a statistical analyses and we are trying to find out the linear regression model the story does not always go in the nice way there can be different types of problems.

And our objective is that that we try to start the linear regression analyses using all the possible variables and we try to observe from the statistical output that what can be different reasons are what are the different types of basic fundamentals which are indicating that the linear regression analyses outcome as some problems based on that once we have a clear understanding of the basic concepts we can diagnose more clearly.

And more specifically about the presence of a particular type of problem in the dataset and then we have to accordingly move forward. One thing I would like to inform all of you that in this course we have considered only the multiple linear regression model under some types of standard assumption. Once these assumptions are not fulfilled we have developed the diagnostics, but we have not discussed, what are the different solutions?

Right, so that part we not covered I am sure that in case if you first try to understand this basic fundamentals possibly in the next course or even yourself you can study from the books and can really understand that if you are facing a particular type of problem like heteroskedasticity or say autocorrelation or the independent variables are not independent that is the problem of multicollinearity, how to obtain a good solution.

So the objective of this example was to show you that the linear regression model is not usually obtain in a single step and we have to revise the model again and again till we get a good model and usually it happens that if we are trying to do it honestly finally we have a good model. So we will see you in the next lecture and till then good bye.