**Data Science for Engineers**
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 34
Diagnostics to Improve Linear Model Fit

Good morning. In the previous lecture we saw different measures that we can use to assess whether the linear model that we have fitted is good or not. For example, we can use the r squared value, and if we find that the r squared value is close to + 1, we can maybe accept the fact that the linear model is good. We can also check based on the f statistic whether the reduced model is better than the model with the slope parameter. So, if we reject the null hypothesis, there again we may conclude that the linear model is acceptable.

We can also do this by testing the significance of the slope parameter. We can look at the confidence interval for the slope parameter and if that does not include 0, then maybe we can accept a linear model. But all of these measures are not sufficient they only provide a initial indicator. I will show you some data set which shows that that these measures are not completely sufficient to accept a linear model. We will use other diagnostic measure to conclusively accept or reject a linear model fit. So, let us look at a data set provided by Anscombe.

(Refer Slide Time: 01:38)
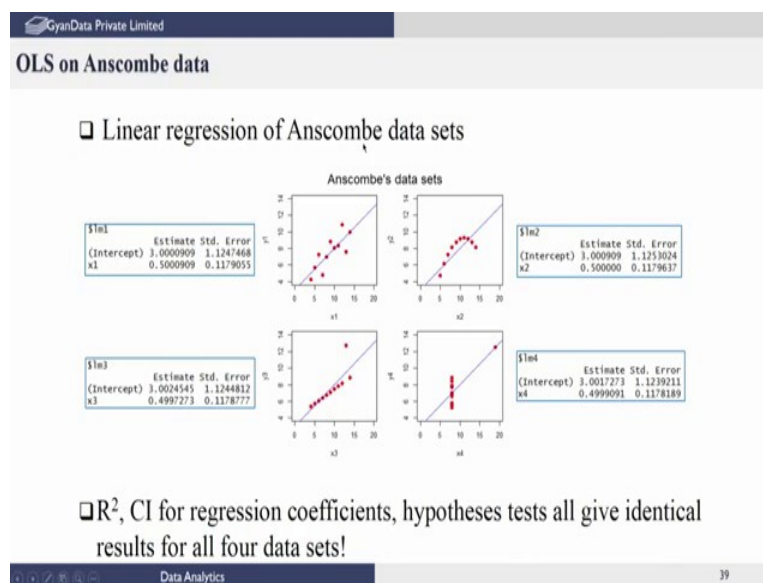
We have seen this before in the when we analyzed statistical measures in one of the lectures, the Anscombe data set is consists of 4 data sets. Each one of them having 11 data points, x versus y they are synthetically constructed to illustrate the point. For example, these 4 data sets are plotted x versus y, the scatter plot is given, first data set here second third and fourth. And in all of these if you actually look at it look at the scatter plot we may say that look a linear model is adequate for the first data set, and perhaps for the third data set, but the second data set indicates that the linear model may not be a good choice. A non-linear model or a quadratic model may be a better fit. The last data set is a very poorly a designed data set. You can see that the experiment is conducted only at 2 distinct values of x, you have one value of x here for which you have 10 experiments conducted you have got 10 different y values for the same x. And then you have one more experimental observation at a different value of x.

So, you should in this case you should not attempt to fit a linear model with the data. Instead, you should ask the experimenter to go and collect data different values of x, then come back and try to check whether that is valid. Unfortunately, when we actually apply linear regression to these data sets, and then find the slope and the intercept parameter we find that in all 4 cases we get the same intercept value of 3, you can see that all 4 data sets you get a value of 3, and you also get the same slope parameter which is point 5 in all 4 cases.

So, the regression model if you are fit to any of these data 4 data sets you will get the same estimate of the intercept and slope. Furthermore, you get the same standard error of fit which is 1.12 for intercept and point one for the slope. And if you run a confidence interval for the slope parameter, you may end up accepting that this slope is acceptable for all 4 cases. And you may conclude incorrectly conclude that the linear model is adequate. You can actually run the r squared value it will be the same for all 4 data sets. You can run the a hypothesis test whether a reduced model is acceptable compared to a model with the slope parameter, again you will reject the null hypothesis using the f statistic. And you may conclude for all 4 cases, you will get the same identical result that a linear model is a good fit. Clearly it is not so.

One can of course, do scatter plots and try to judge it in this particular case. Because it is a univariate example, but when you have many independent variables, then you have to examine several such scatter plots and that may not be very easy.

So, if you assume there are 100 independent variables, you have to examine 100 such plots of y versus x. And it may not be possible for you to make a visual conclusion from that. So, we will use other kinds

of plots called residual plots, which will enable us to do this whether it is a univariate regression problem or a multivariate regression problem, we will see what these are.

(Refer Slide Time: 04:57)
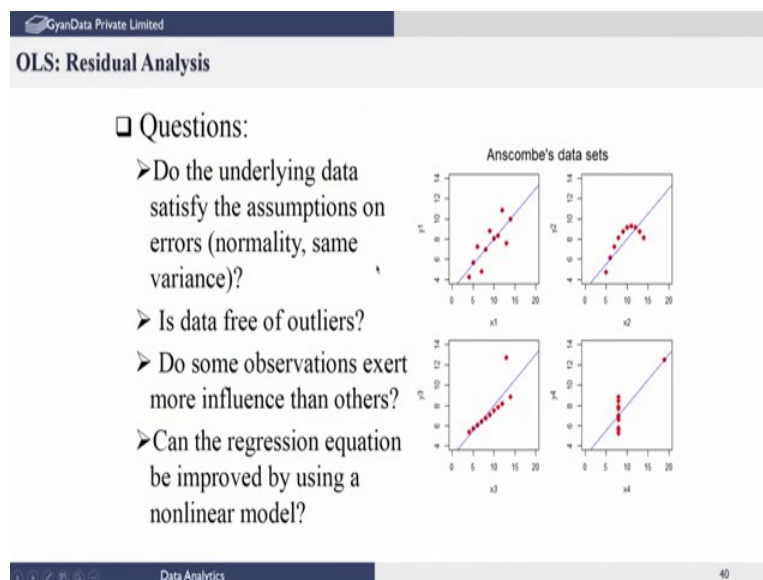
So, the main questions that we are trying to ask now whether a linear model is adequate. We have some measures we have seen, but they are not adequate, we will use additional things, and when we did the linear regression, we did make additional assumptions although they were not been stated explicitly, we assume that the errors that corrupt the dependent variables are normally distributed and they have identical variance. Only under this these assumptions can you use a least squares method to perform linear regression. That you can at least prove that the least squares method has some nice properties.

So, we do not know whether this is true and we have to verify whether the errors are normally distributed and have equal variance. We also may have a problem of data containing outliers which we may have to remove, and that also we have to solve. Additional questions may be that some observations may have unduly high influence than others and we want to identify such points and perhaps remove them or at least be aware of this. And lastly of course, a linear model maybe inner equates. So, we have to try and t a non-linear model. So, I am going to only address the first 2 questions, whether the errors are normally distributed, whether they have equal variance and whether there are outliers in the data. These 2 things we will address using residual plots. So, let us do this illustrate with the anscombe data set and also other data set.

So, one way of assessing, whether they are outliers or whether linear model is adequate or not is using what we call residual plots. And let us see what these residuals are. By definition a residual is the difference between the measured value of the dependent variable - the predicted value of the dependent variable for each sample. So, yi represents the measured value, ŷi is the predicted value using the linear regression model that we are fitted. So, that difference is designated as ei, and it is called the residual. And that is nothing but the vertical distance between the fitted line and the observation point. Now we can try to compute the statistical properties of these residuals, and we will be able to show that the variance of these residuals are not all identical, even though we started with the assumption that all errors corrupting the measured values have the same variance.

But the residual which is a result of the fit will not have the same variance, for all data points. In fact, you can show that the variance of the ith sample is σ^2 which represents the error in the measured value of the dependent variable multiplied by 1 - pii; where pii is defined by this. Notice that pii depends on the ith sample, numerator depends on the ith sample, therefore p ii depends on the ith sample and varies with sample to sample.

So, the variance of the residual will not be identical for all examples, it is given by this quantity. We also can show that the residuals are not independent even though we assume that the errors corrupting the measurements are all independent. The residuals in the samples are not independent and they have a correlation covariance and that covariance can be shown to be given by this quantity. The reason for the variance not being identical of the residuals or them

being correlated is because you notice that this ŷi they have actually have here is a result of the regression it depends on all variables all measurements. It is not dependent only on the ith measurement. This predicted value is a function of all the observations, and that because of that it introduces a correlations between the different residuals. And also, imparts different variants to different residuals. And so, having derived this, notice that even if we do not have a priori knowledge of σ square which is the variance of error in the measurements we can estimate this quantity.

We have already seen this in the previous lecture, we can replace this by s s e by n - 2, which is an estimate of this σ^2 and substitute this to get an estimated variance of each residual.

(Refer Slide Time: 09:34)

We will standardize these residuals, where what we mean by standardization is to divide the residual by it is standard deviation estimated standard deviation. All of this can be computed from the data, and therefore, you get for each sample a standardized residual after performing the linear fit which is given by this quantity. Now you can also show that this particular quantity the standardized residual will have a t distribution with n - 2 degrees of freedom. Now the these statistical properties allow you to now perform test on the residuals, which will what we will use, to identify outliers and also test whether there is set the variances in the different measurements are identical or not.

So, we will plot the residual what we call a residual plots will we plot the residuals with respect to the predicted or fitted value of the dependent variable remember there is. Only one dependent variable even if there are multiple independent variables we have only one dependent variable, we can plot the residuals with respect to the predicted value of the dependent variable. And the predicted values you obtain after the regression remember.

So, this is called the residual plot what is called the residual versus the fitted or predicted value and this plot is very useful in testing the validity of the linear model, in determining whether the errors are normally distributed, assumptions on errors are ok and whether the variances of all errors are identical or not which is called the homoscedastic case, which means the errors in all measured values are identical or the variance of the error in different measured values are non-identical which is called the heteroscedastic case or heteroscedastic error. So, let us see how each of these how the plot looks for each of these cases.
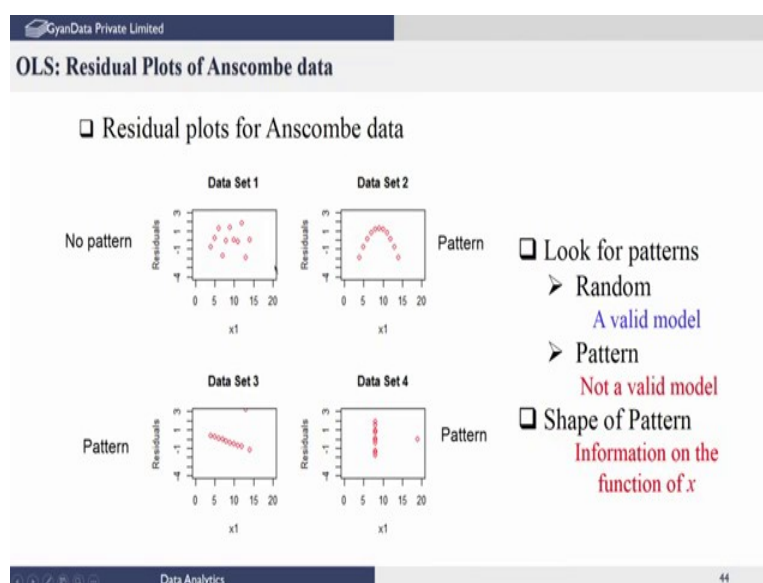
Now, let plot the residual plot for the 4 data sets provided by Anscombe. Notice that we have done the regression model regression model we computed all these parameters r squared confidence interval they all turned out to be identical they gave us no clues, whether the linear model is good for all 4 data sets or not. Basically they say they would say that the linear model is adequate.

But when we do the residual plot here we have plotted the residual versus the I have plotted with respect to the independent variable, but because it is a univariate case, we have plotted with respect to the independent variable. But technically you should plot the residual with respect to the predicted value of the dependent variable, remember because we presume that the predicted variable is linearly dependent on x, in this case it may not matter the pattern will look the same you can try it out for yourself if you plot the residual with respect to the predicted value of the variable dependent variable. Then you will get this kind of pattern of the residuals for the 4 data sets. The first data set if you look at it exhibits no pattern. The residuals seems to be randomly distributed between this case between - 3 and + 3 and whereas for the second data set there is a distinct pattern. The residuals look like a quadratic like a parabola. And so, therefore, there exists a pattern in the data set 2. For the third data set basically, you can say that there is no pattern, except that are constant more or less linear are constant, there seems to be a small bias because of the slope left in the residuals. Data set 4 as we saw before is a poorly designed experimental data set. All the y values are obtained at a single x value and that is what the residuals are also showing. The 10 of the data points obtained at the same x value are showing different residuals and the one single residual at a different x value showing something.

So, from this you cannot judge anything. All you can say is that the experimental data set is very poorly designed, and you need to get back to the experimenter and ask him to provide a different data set. Now based on this we can safely conclude that data set one clearly a linear model is adequate. All the measures, previous measures also, were satisfied. And now the residual plot also shows a random pattern which means a random or what we call no pattern, then linear model is adequate. Whereas, for data set 2 by look at the residual plot we can conclude that a linear model is inadequate should not be used for this data set.

For the third data set however, we know there is one data point that is lying far away, and perhaps that is the one that is causing all of this slightly a linear pattern here. And if we remove this outlier and retry it maybe this resolve this problem will get resolved, and linear model may be adequate for data set 3. For data set 4 again there is a distinct constant pattern and therefore, we can conclude that linear model should not be used.

In fact, no model should be used between x and y, because y does not seem to be dependent on x here. So, the residual plot clearly gives the game away, and it should be used along with other measures in order to finally, conclude that the linear model that were fitted for the data is acceptable or not. So, in this case data set one certainly will accept data set 2 3 we will have to do further analysis. But for 2 and 4 we will completely reject the linear model.

(Refer Slide Time: 14:59)

Now, the test for normality can also be done using the residuals. We have already seen the prob even we did statistical analysis the notion of a probability plot. Where we plot the sample quantiles, against the quantiles from the distribution with which we want to compare.

So, if we want to compare, whether a set of given a given set of data. Follows a certain distribution then we plot the sample quantiles from the quantiles drawn for that particular distribution against which we want to test this sample. So, in this case we want to test, whether residuals that we have the standardized residuals, come from a standard normal distribution. And therefore, we will take the quantiles from the standard normal distribution and plot it. Just to recap what do we mean by quantile it is a percentile data value below which a certain percentage of data lies. For example, if you want to find given a data set, what is the value below which 10 percent of the data lies, maybe - 1.28 here we are given, which means 10 percent of the samples lie below - 1.28. 20 percent of the samples lie below - 0.84 and so on and so forth.

We have computed this, this we can plot against the standard normal values 10 percent value where the probability, between - ∞ and the value is 10 percent. And the value between - ∞ and that value should be 20 percent and so on so forth. Those represents the x values corresponding to these probabilities and we can use that plot it and then before completing this contents we have arranged the data.

So, we have seen this before I have just only recapped this. And we can use what is called Q-Q norm function in r to actually do it if you give the data set x and ask you to do a probability plot Q-Q norm will do this for you directly in r.

(Refer Slide Time: 17:03)
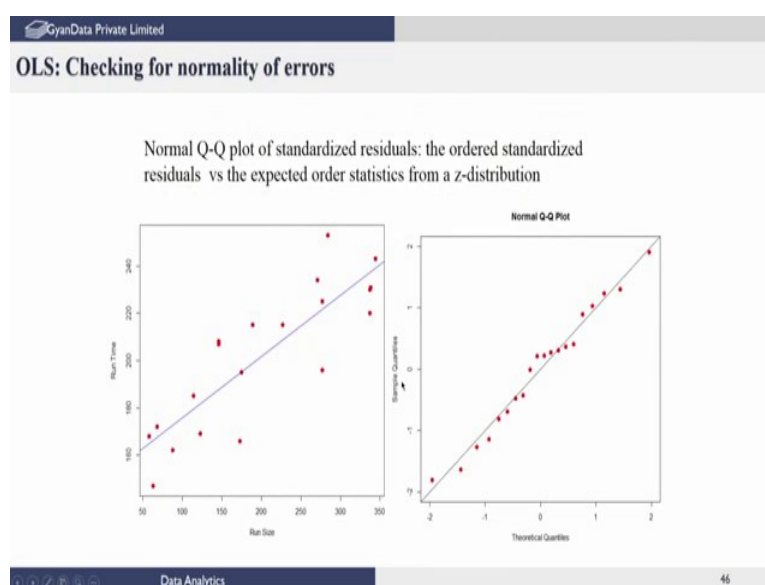
So, this is a sample Q-Q plot I have taken for some arbitrary random data set samples drawn from the standard normal distribution and you can see that if you do the normal Q-Q plot for the residuals after fitting the regression line, it seems to closely follow the 45-degree line. So, the theoretical quantiles computed from the standard normal distribution and the quantiles computed from the sample residuals, standardized residuals, the fall on the 45-degree line. And therefore, in this case we can safely conclude that the errors in the data come from a standard normal distribution.

So, a Q-Q plot if this thing is in does not happen. If we find the significant deviation of this quantiles from the 45-degree line, then a normal distribution assumption is incorrect. Which means we have to modify the a least square subjective function to actually accommodate this. It may not have it may or may not have a significant effect on the regression coefficient, but there are ways of dealing with it which I will not go into.

(Refer Slide Time: 18:13)
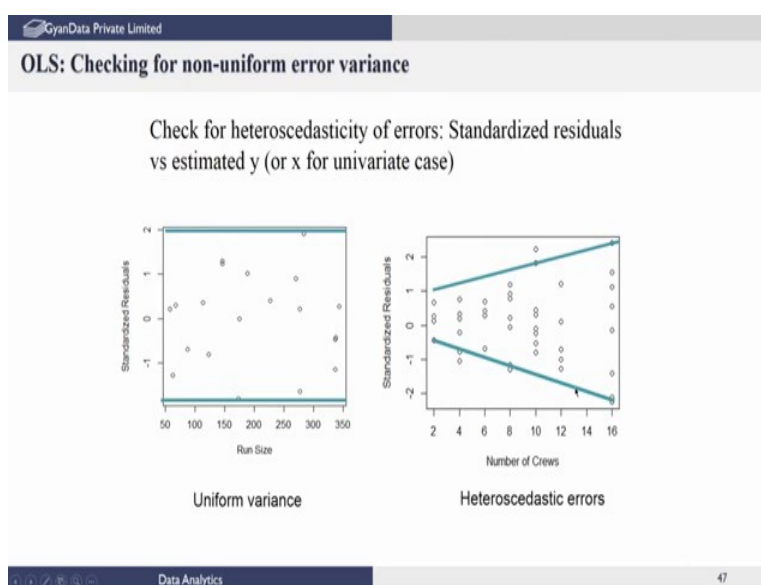
A third thing that we need to test is whether the residual variances I am sorry the error variances in the data are having a uniform variance or I have different variances for different samples, and again here what we do is look at the residual plot and standardized residual versus the predicted values what you have to plot and if you do and look at this thing. It seems to be that the there is no particular trend in the residuals. For example, in the right-hand side we find that the residuals close to when the number of values is - 2 is 2 is spread is very small.

Whereas, when the number of crews is 16 the spread is very high. So, the spread increases or looks like a funnel when we actually look at

the residuals. Whereas, such a effect is not found on the data set corresponding to the left-hand side thing. So, here I have plotted the standardized residual for 2 different data sets. Just to illustrate the type of figures you might get. If you get a figure such as in the left then we can safely conclude that the errors in different measurements have the same variance, whereas if you have a funnel type of effect then you know that the errors, where a variances increases as the value increases. So, it depends on the value itself, which implies that you cannot use a standard least squares method, you should use a weighted least squares method.

So, data points which are corresponding to these 4 should be given more weight and data points corresponding to this should be given less weight and we call that a weighted least squares. That is the way we have to deal with what we call heteroscedastic errors of the sky. Again, I am not going to go into the whole thing I just want to illustrate, that first the residual plots are used in order to verify the assumptions and if the assumptions are not valid then we have correction mechanisms to modify our regression procedure. But linear this does not indicate a linear model is not adequate, the linear model adequacy test is basically based on the pattern if there is no pattern in the residuals you can go ahead and assume that the linear model t is adequate as long as the other measures are also satisfactory. But here it is related to the error variances, and in this case, we only modify the linear regression method, and we still go with a linear model for this for these case such as the one shown on the right.

(Refer Slide Time: 21:01)

The last thing that we need to do is also clean out the data, we do not want to use data that have got large errors what we call outliers, points which do not conform to the pattern that is found in the bulk of the data. And the outliers can be easily identified using hypothesis test of the residual for each sample. We have actually found the residual for example, we have actually found the standardized residuals. So, the standardized residual roughly follow we know it follows a t distribution. But we can for large enough number of samples, we can assume that it follows a normal distribution. So, if I use a 5 percent level of significance we can run a test, hypothesis test for each sample residual, and if the residual lies outside - 2 to + 2, we can conclude that the sample is an outlier.

So, for each sample, we test whether the sample standardized residual lies outside of this interval and if it lies outside this interval, we can conclude that that particular sample maybe an outlier and remove it from our data set. The only thing when we do out lie detection is this, it may turn out that we do that first time we fit a regression, and do an outlier detection we may find several residuals. Lying outside the confidence interval - 2 to + 2 -95 percent confidence interval in which case we do not throw all the samples out at the same time. We only throw out the one that is most o ending. Which is we identify the outlier that corresponds to the sample with the largest standard standardized residual magnitude; which of which is furthest away from - 2 or + 2. That is the one we will take and remove it.

Once we remove that we again run a regression on the remaining samples, and again run this outlier detection test. So, we do remove only one outlier at a time. The reason for this is, when we perform an outlier detection we should be aware that a single outlier can smear affect the residuals of other samples because of our regression parameters are obtained from all the data points. Therefore, even a single outlier can cause other outliers to fall outside the confidence interval. Therefore, we do not want hastily conclude that all residuals falling outside the confidence interval are outliers. Only the one that has the maximum magnitude we actually take it out. And then we redo this so, that one at a time we do it will be a safe way of performing outlier detection.
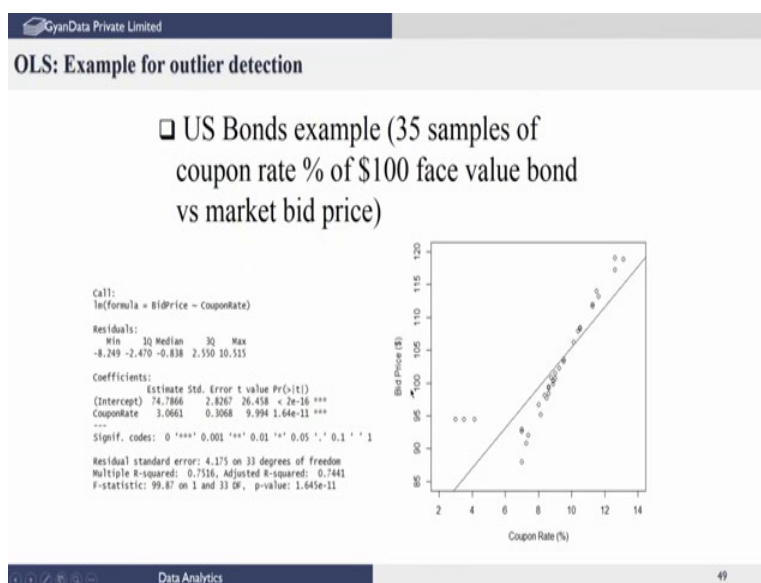
Again, we will illustrate this with an example. Here is us bonds example which consists of 35 samples, us bonds whose face value is 100 dollars, and it is a guaranteed interest rate is provided for each of these bonds depending on when they were released and so on, and that are different bonds with different interest rates. But these are also traded in the market, and the selling price in the market or bid price would be different depending on the kind of interest rate they attract. So, you would presume that the bond which has a higher interest rate would have a higher market price.

So, there might be a linear correlation or linear relation between the market price and the interest rate for that bond. So, here there are 35 samples that are obtained from nothing. These data sets are standard data sets that you can actually download from the net, if you just search for it, you will get it just like the Anscombe data set, and the and the what you call computer repair time data set that I have been using in the previous lectures. If you perform a regression, and you get a t of this kind. So, it shows that the linear t seems to be adequate you can run the r command lm and you will find that the intercept is 74.8 and the slope is 3.6, and standard errors given and clearly the what you called the p value is very, very low; which means that you will not reject the significance that is the intercept is significant and the slope is also significant, they are not close to 0. You can of course, compute confidence interval and come to the same judgment.

You can run an f test, here also it says the p value of the f test is -11, which means you will reject the null hypothesis. And conclude that a full model is adequate which means the slope is important here. So, the r value seems to be reasonably good 0.75. And so, we can say the

initial indicators are that a linear model is adequate. Now let us go ahead and do the residual analysis for this.

(Refer Slide Time: 25:44)
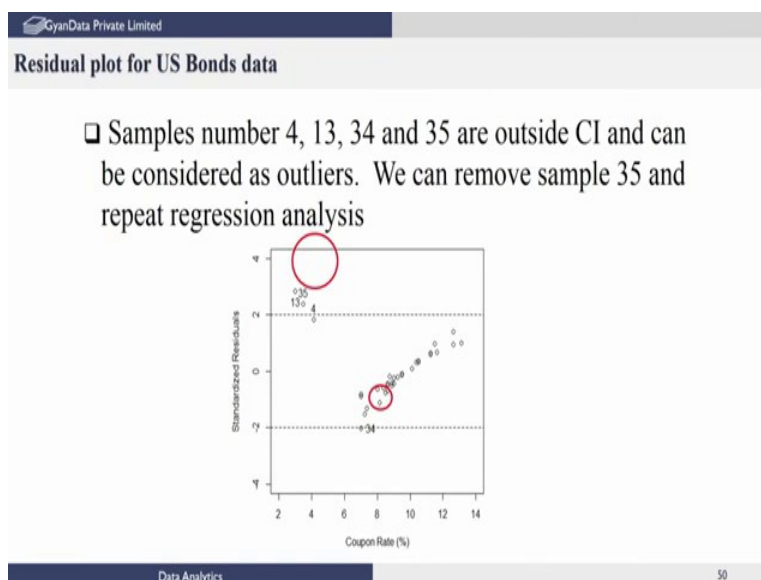
We perform a residual plot for standardized residual plot, and we find that except for these 4 points, 35 sample number 35 , sample number 13, sample number 34 seems to be outside of the + - 2 confidence interval, and they maybe you may conclude that these are out outliers. While the others are within the bond and they are de nitely not out outliers. There seems to be some kind of a pattern, as the coupon that increases the standardized residuals increase. So, maybe as there is a certain amount of non-linearity in the model, but let us remove these outliers before making a final conclusion.

So, we can remove all these 4 outliers at the same time, if you want as I said that is not a good idea perhaps we should remove only sample number 35 which has furthest away from the boundary with the residual with the largest magnitude. And redo this because of lack of time, I have just removed all 4 at the same time and then done the analysis. My suggestion is you do one at a time and then repeat this exercise for yourself. Here we have removed these 4 samples 4 13 35 and thing and run the regression analysis.

(Refer Slide Time: 27:02)
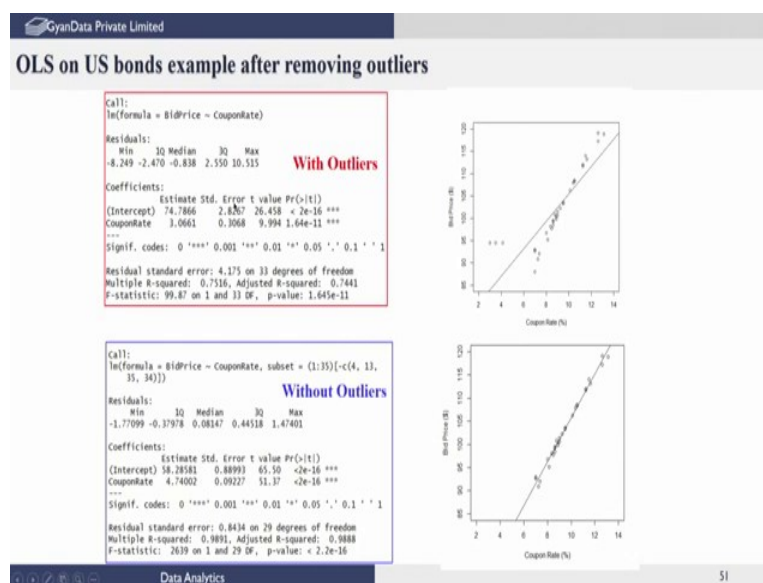
Again, you can see that the regression analysis, maintain retaining all the samples is shown on the right-hand side of the plot. And their corresponding intercept coupon the slope as well as the f test statistic and so on r squared value is shown here. And once we remove these 4 samples which we outliers and then rerun it, now the fit seems to be much better. It is also seen on the left-hand side that the t is much better you can see that the r squared value has gone up to 0.99 from 0.775. The again the test on the intercept and the coupon rate or slope shows that that they are significant and therefore, you should not assume that they are close to 0.

It also shows that the f statistic has also a low p value which means, you raise the null hypothesis that reduce model is adequate which means the linear model with the slope parameter is a much better fit. So, all of this indicator show seem to indicate show that a linear model is adequate and the t seems to be good. But we should do a residual plot again with this data. And if that actually shows no pattern we can actually stop there we can say that are no outliers. And therefore, we can conclude that regression model that we have fitted for this data is a reasonably good one.

Next class will see how to actually extend all of these ideas to the multiple linear regression which consist of many independent variables and one dependent variable.

Thank you.