Python for Data Science
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

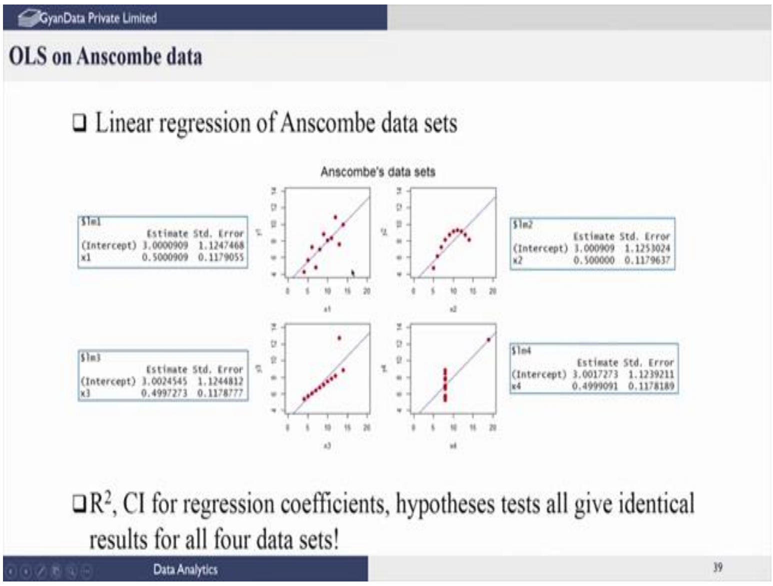
Lecture – 29
Diagnostics to Improve Linear Model Fit

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 plus one 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 an 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)



GyanData Private Limited

OLS on Anscombe data

❑ Linear regression of Anscombe data sets

Anscombe's data sets

```
$lm1
            Estimate Std. Error
(Intercept) 3.0000909  1.1247468
x1          0.5000909  0.1179055
```

```
$lm2
            Estimate Std. Error
(Intercept) 3.000909  1.1253024
x2          0.500000  0.1179637
```

```
$lm3
            Estimate Std. Error
(Intercept) 3.0024545  1.1244812
x3          0.4997273  0.1178777
```

```
$lm4
            Estimate Std. Error
(Intercept) 3.0017273  1.1239211
x4          0.4999091  0.1178189
```

❑ R^2, CI for regression coefficients, hypotheses tests all give identical results for all four data sets!

We have seen this before in the when we analyzed statistical measures in one of the lectures. The Anscombe data set is consists of four data sets each one of them having 11 data points, x versus y. They are synthetically constructed to illustrate the point. For example, these four 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 long linear model or a quadratic model may be a better fit. The last dataset is a very poorly designed data set, you can see that the experiment is conducted only at two 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.

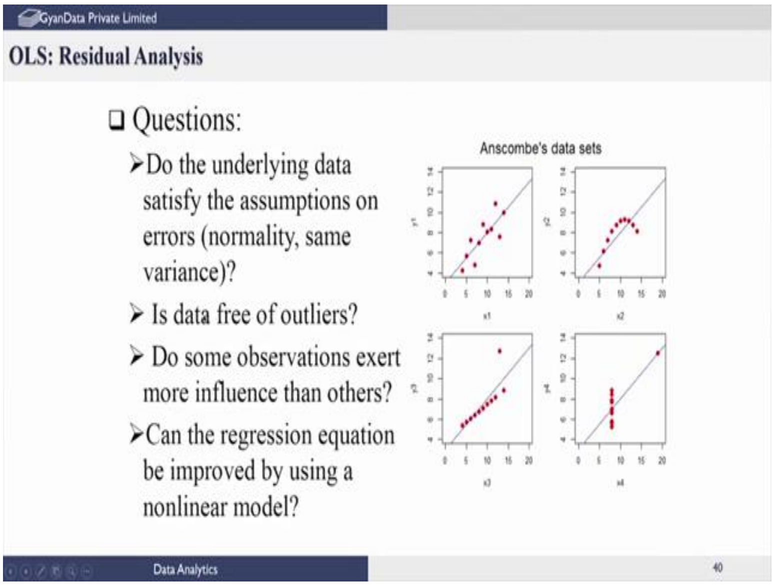
So, you should in this case you should not attempt to fit a linear model with the data, instead you should ask the experimenter to go and collect data at different values of x then come back and try to check whether that is valid. Unfortunately, when we actually apply linear regression to these data sets and then find the slope and the intercept parameter we find that in all four cases we get the same intercept value of 3 you can see that all four data sets you get a value of 3 and you also get the same slope parameter which is 0.5 in all four cases.

So, the regression model if you fit to any of these data four data sets you will get the same estimate of the intercept and slope. Furthermore, you get the same standard error of it 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 four 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 four 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 a null hypothesis using the F statistic and you may conclude for all four cases you 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)



GyanData Private Limited

OLS: Residual Analysis

❑ Questions:

- Do the underlying data satisfy the assumptions on errors (normality, same variance)?
- Is data free of outliers?
- Do some observations exert more influence than others?
- Can the regression equation be improved by using a nonlinear model?

Data Analytics

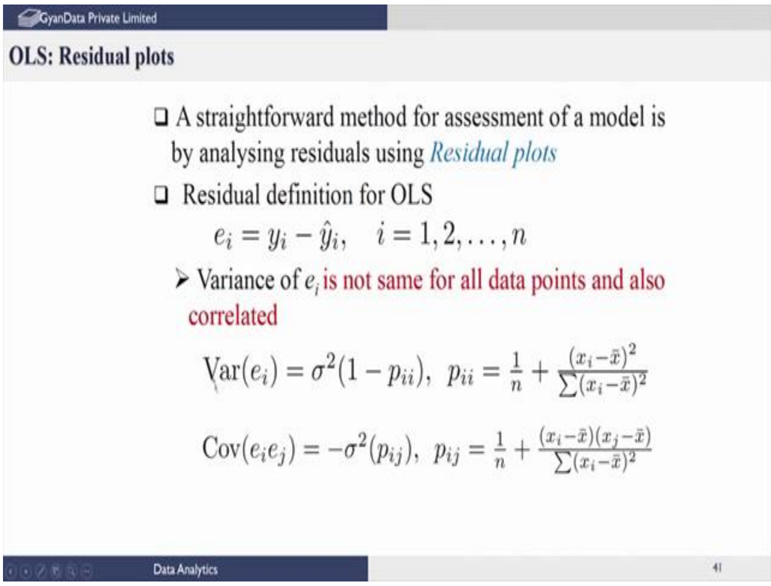
40

So, the main questions that we are trying to ask now whether a linear model is adequate? We have has some measures we have seen, but they are not adequate. We will use additional things and when we needed linear regression we did make additional assumptions although they have not been stated explicitly. We assumed that the errors that corrupt the independent 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 way 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 may be inadequate so, we have to try and fit a non-linear model.

So, I am going to only address the first two questions; whether the errors are normally distributed? Whether they have equal variance and whether there are outliers in the data? These two things we will address using residual plots. So, let us do this illustrate with the Anscombe data set and also other data set.

(Refer Slide Time: 06:27)



GyanData Private Limited

OLS: Residual plots

- ❑ A straightforward method for assessment of a model is by analysing residuals using *Residual plots*
- ❑ Residual definition for OLS

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

➢ Variance of e_i is not same for all data points and also correlated

$$\text{Var}(e_i) = \sigma^2(1 - p_{ii}), \; p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

$$\text{Cov}(e_i e_j) = -\sigma^2(p_{ij}), \; p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum(x_i - \bar{x})^2}$$

Data Analytics

41

So, one way of assessing whether there are outliers or whether linear model is adequate or not is using what we call the 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, y_i represents the measured value, $\hat{y}_i$ is the predicted value using the linear regression model that we have fitted. So, that difference is designated as e_i and it is called the residual and that is nothing, but the vertical distance between the fitted line and the observation point.

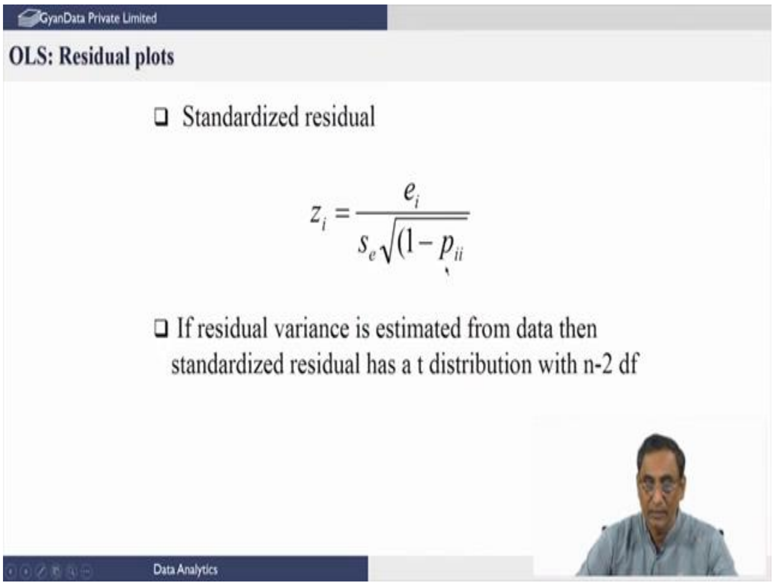
Now, we can try to compute the statistical properties of these residuals and we will be able to show that the variance of this these residuals are not all identical even though we started with the assumption that all errors corrupting the measured values of the same variance, but if residual which is 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 i-th sample is sigma squared which represents the error in the measured value of the dependent variable multiplied by one-p ii, where p ii is divide by this.

Notice that p_{ii} depends on the i-th sample, numerator depends on the i-th sample therefore, p_{ii} depends on the i-th sampled and varies with sample to sample. So, the variance of the residual will not be identical for all samples, 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 their them being correlated is because you notice that this $\widehat{y_\iota}$ we have actually have here is a result of the regression. It depends on all variables, all measurements; it is not depend only on the i-th measurement. This predicted value is a function of all the observations and that because of that it introduces a correlations between the different a residuals and also imparts different variance to different residuals.

And so, having derived this notice that even if we do not have a priori knowledge of sigma square which is the variance of error in the measurements we have we can estimate this quantity. We have already seen this in the previous lecture. We can replace this by SSE by n-2 which is an estimate of the sigma square and substitute this to get an estimated variance of each residual.

(Refer Slide Time: 09:34)



GyanData Private Limited

OLS: Residual plots

❑ Standardized residual

$$z_i = \frac{e_i}{s_e\sqrt{(1-p_{ii}}}$$

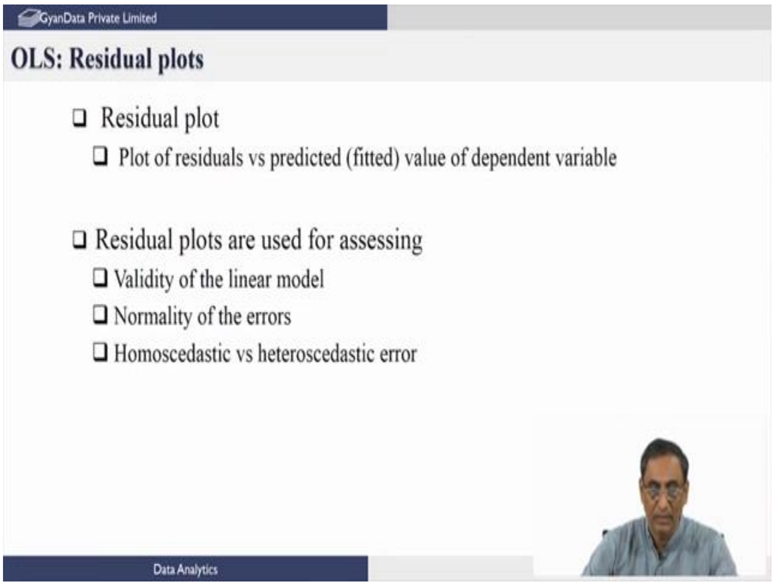
❑ If residual variance is estimated from data then standardized residual has a t distribution with n-2 df

Data Analytics

We will standardize these residuals, where what we mean by standardization is to divide the residual by its standard deviation estimated standard deviation ok. 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, what these statistical properties allow you to know perform test on the residuals which what we will use to identify outliers and also test whether there is set the variances in the different measurements are identical or not.

(Refer Slide Time: 10:20)



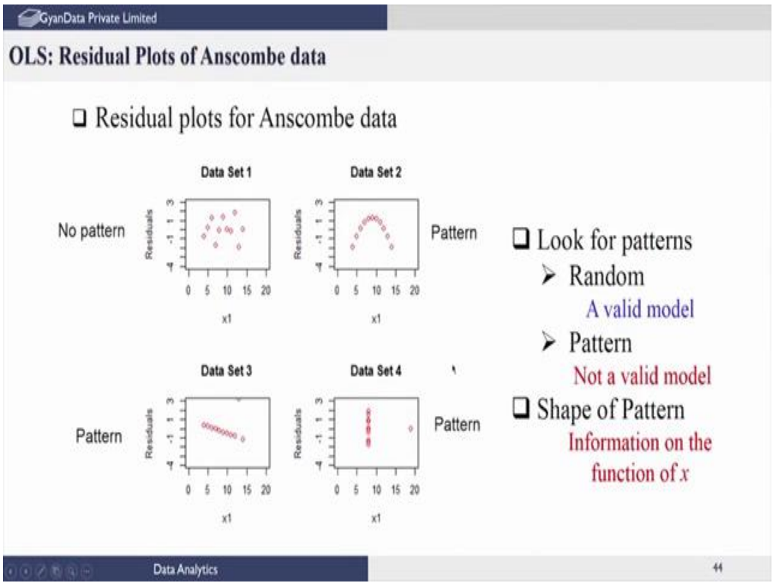
OLS: Residual plots

- ❑ Residual plot
  - ❑ Plot of residuals vs predicted (fitted) value of dependent variable
- ❑ Residual plots are used for assessing
  - ❑ Validity of the linear model
  - ❑ Normality of the errors
  - ❑ Homoscedastic vs heteroscedastic error

So, we will plot the residual what we call residual plots we will 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 plug the residuals with respect to the predicted value of the dependent variable and the predicted values will 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 and whether the variances of all errors are identical or not which is called a homoscedasticity case which means the errors and 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.

(Refer Slide Time: 11:28)



GyanData Private Limited

OLS: Residual Plots of Anscombe data

❑ Residual plots for Anscombe data

❑ Look for patterns

- ➢ Random

  A valid model

- ➢ Pattern

  Not a valid model

❑ Shape of Pattern

Information on the function of x

Data Analytics

44

Now, let us plot the residual plots for the four 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 four 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 are plotted the residual versus the I have plotted with respect to the independent variable.

But, because it is a univariate case we are 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, if 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 dependent variable then you will get this kind of pattern of the residuals for the four data sets.

The first dataset if you look at it exhibits no pattern. The residuals seem to be randomly distributed between this case between-3 and plus 3 and whereas, for the second dataset 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 our constant, more or less linear or constant. There seems to be a small bias because of the slope left in the residuals.

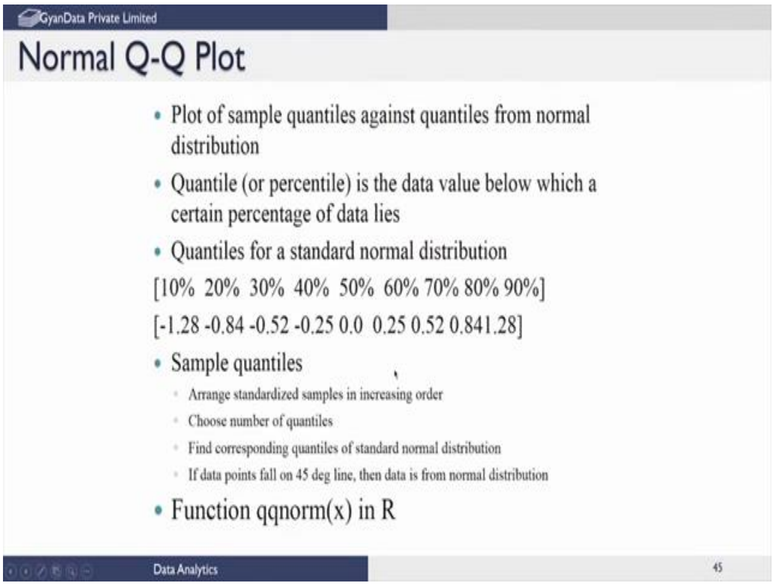
Data set 4 as we saw before, is a poorly designed experimental data set. All the y values are obtained at a single x value and that is what the residuals are also showing the 10 of the data points obtained that the same x x value or showing different – different residuals and the one single residual 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 we 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 1 clearly linear model is adequate all the measures previous measures also were satisfied and now the residual plot also shows a random pattern which means or random or what we call no pattern, then linear model is adequate whereas, for data set 2 by looking 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 linear pattern here. And, if we remove this outlier and retry it maybe this will this resolved this problem will get resolved in 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 a linear model should not be used. In fact, no model should be used between x and y because y it 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 1 certainly will accept, data set 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)



GyanData Private Limited

Normal Q-Q Plot

- Plot of sample quantiles against quantiles from normal distribution
- Quantile (or percentile) is the data value below which a certain percentage of data lies
- Quantiles for a standard normal distribution

[10% 20% 30% 40% 50% 60% 70% 80% 90%]

[-1.28 -0.84 -0.52 -0.25 0.0 0.25 0.52 0.841.28]

- Sample quantiles
  - Arrange standardized samples in increasing order
  - Choose number of quantiles
  - Find corresponding quantiles of standard normal distribution
  - If data points fall on 45 deg line, then data is from normal distribution
- Function qqnorm(x) in R

Data Analytics 45

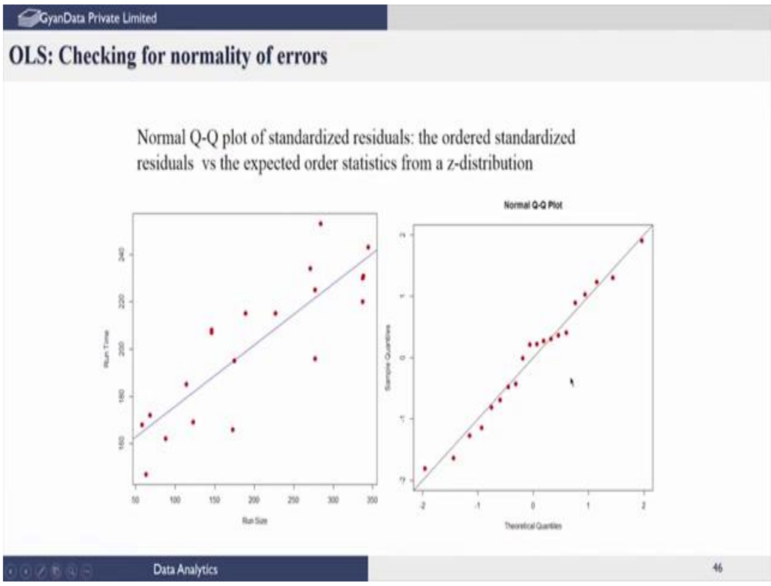
Now, the test for normality can also be done using the residuals. We have already seen the when we did statistical analysis the notion of a probability plot where we plot the sample quantiles against the quantiles from the distribution with which we want to compare. So, if we want to compare whether 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 the 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 a 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 probably where the probability between-infinity and the value is 10 percent and the value between-infinity 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 plotted and then of course, before computing these contexts we have arranged the data.

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

(Refer Slide Time: 17:02)



GyanData Private Limited

OLS: Checking for normality of errors

Normal Q-Q plot of standardized residuals: the ordered standardized residuals vs the expected order statistics from a z-distribution

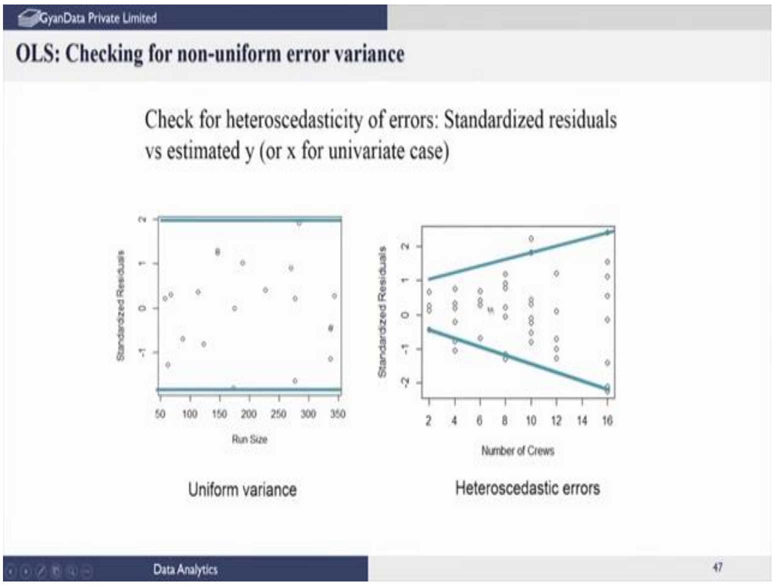
Data Analytics

46

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 follow on the follow 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 least squares objective function to actually accommodate this. It may not have 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: 18:12)



GyanData Private Limited

OLS: Checking for non-uniform error variance

Check for heteroscedasticity of errors: Standardized residuals vs estimated y (or x for univariate case)

Uniform variance

Heteroscedastic errors

Data Analytics

47

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 residuals 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.

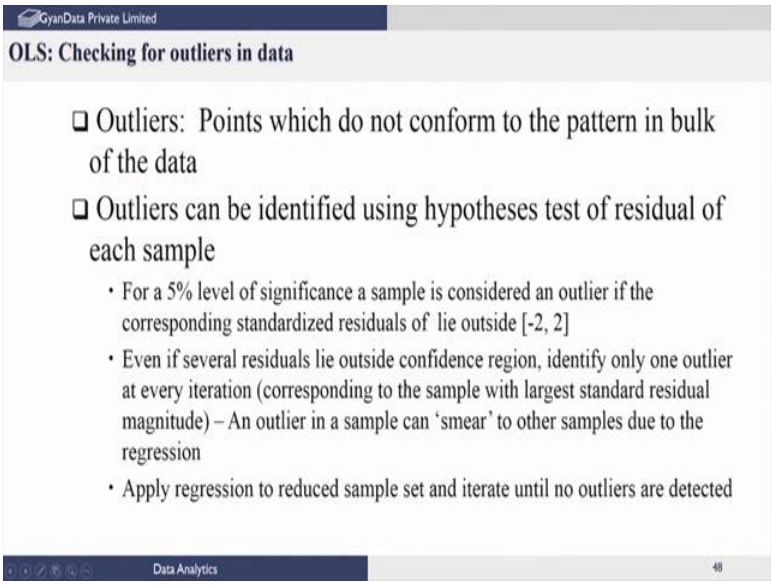
For example, in the right hand side we find that the residuals close to when the number of values is-2 and 2 is spread is very small whereas, when the number of crews is 16 the spread is very high. So, the spread increases or looks like a funnel when we actually look at the residuals whereas, such a effect is not found on the data set corresponding to the left hand side thing.

So, here I have plotted the standardized residual for two 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 will 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 this these four should be given more weight and data points corresponding to this will be given less weight and we call that a weighted least squares that is the way we have to deal with what we call heteroskedastic errors of this guy. Again, I have I am not going to go into the whole thing I have just 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 fit is adequate as long as 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 these cases for this cases such as the one shown on the right.

(Refer Slide Time: 21:01)



GyanData Private Limited

OLS: Checking for outliers in data

- ❑ Outliers: Points which do not conform to the pattern in bulk of the data
- ❑ Outliers can be identified using hypotheses test of residual of each sample
  - For a 5% level of significance a sample is considered an outlier if the corresponding standardized residuals of lie outside [-2, 2]
  - Even if several residuals lie outside confidence region, identify only one outlier at every iteration (corresponding to the sample with largest standard residual magnitude) – An outlier in a sample can 'smear' to other samples due to the regression
  - Apply regression to reduced sample set and iterate until no outliers are detected

Data Analytics

48

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 minus 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 each sample we have actually finally, found a standardized residual. So, the standardized residual roughly follow we know it is follows a T distribution, but we can for large enough number of samples we can assume that it follows a normal distribution.

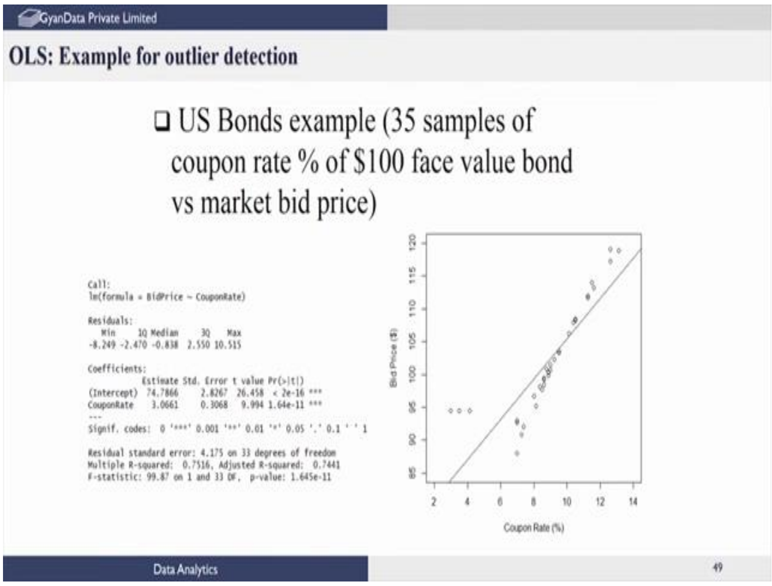
So, if we I use a 5 percent level of significance we can run a test hypothesis test for each sample residual and if the residual lies outside-2 to plus 2, we can conclude that the sample is an outlier. So, for each sample we test whether the sample standardized residual lies outside of this interval and if it is lies outside this interval we can conclude that that particular sample may be an outlier and remove it from a data set.

The only thing when we do outlier 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 plus 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 offending which is we identify the outlier that corresponds to the sample with the largest standard standardized residual magnitude which of which is farthest away from-2 or plus 2 that is the one we 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 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 to hastily conclude that all residuals following outside the confidence interval 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.

(Refer Slide Time: 23:34)



GyanData Private Limited

OLS: Example for outlier detection

❑ US Bonds example (35 samples of coupon rate % of $100 face value bond vs market bid price)

```
Call:
lm(formula = BidPrice ~ CouponRate)

Residuals:
   Min     1Q Median     3Q    Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516, Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Data Analytics 49

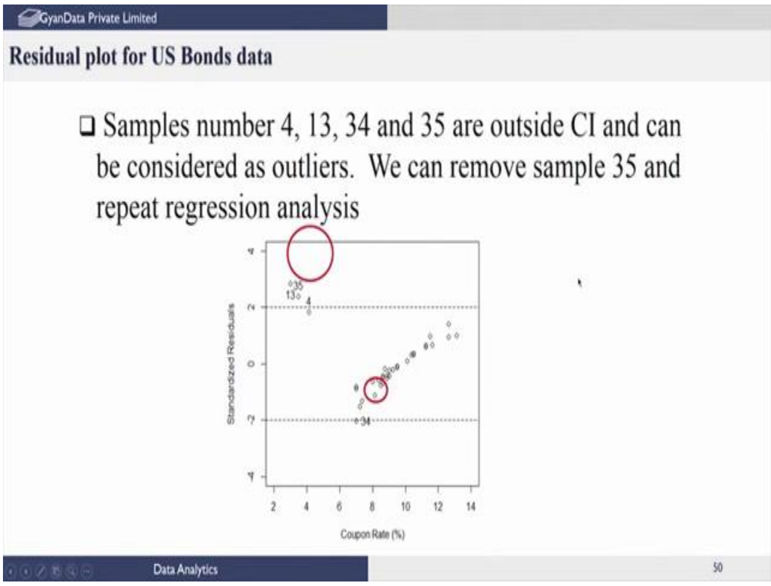
Again we will illustrate this with an example. Here is a US bonds example which consists of 35 samples, US bonds whose face value is 100 dollars and it is a guaranteed interest rate is provided for each of these bonds depending on when they were released and so on and there are different bonds with different interest rates. But, these are also traded in the market and their 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 on linear relation between the market price and the interest rate for that bond. So, here there are 35 samples that are obtained from a thing. These datasets 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 what you called computer repair time data set that I have been using in the previous lectures.

If you perform a regression and you will get a fit of this kind ok. So, it shows that a linear fit 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 call 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 ok. 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: 24:44)



GyanData Private Limited

Residual plot for US Bonds data

❑ Samples number 4, 13, 34 and 35 are outside CI and can be considered as outliers. We can remove sample 35 and repeat regression analysis

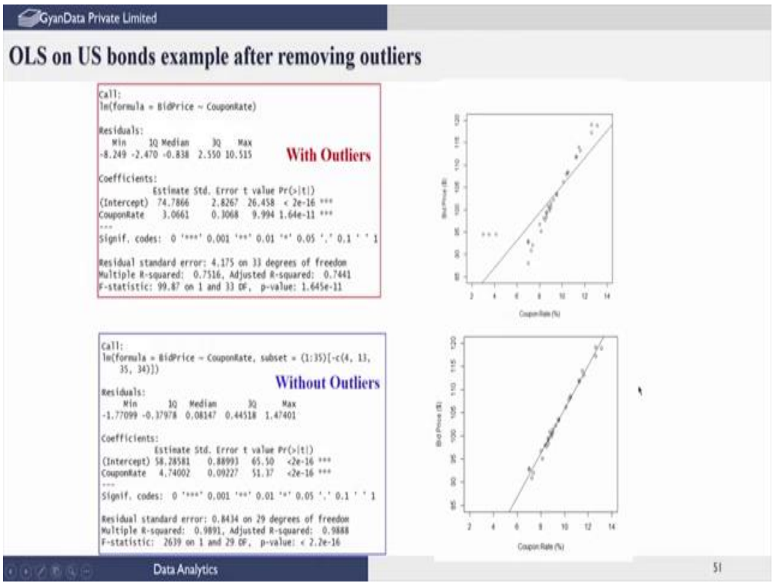
Data Analytics

50

We perform a residual plot for standardized residual plot and we find that except for these four points 35 sample number 35, sample number 13, sample number 34, seems to be outside of the plus-2 confidence interval and they may be you may conclude that these are outliers while the others are within the bound and they are definitely not out outliers. They seem to be some kind of a pattern as the coupon rate 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 four 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 farthest away from the boundary with the residual with the largest magnitude and redo this because of lack of time I have just removed all four 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.

(Refer Slide Time: 27:01)



GyanData Private Limited

OLS on US bonds example after removing outliers

With Outliers

```
Call:
lm(formula = BidPrice ~ CouponRate)

Residuals:
   Min     1Q Median     3Q    Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516, Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Without Outliers

```
Call:
lm(formula = BidPrice ~ CouponRate, subset = (1:35)[-c(4, 13,
    35, 34)])

Residuals:
     Min       1Q   Median       3Q      Max
-1.77099 -0.37978  0.08147  0.44518  1.47401

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.28581    0.88993   65.50   <2e-16 ***
CouponRate   4.74002    0.09227   51.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8434 on 29 degrees of freedom
Multiple R-squared:  0.9891, Adjusted R-squared:  0.9888
F-statistic:  2639 on 1 and 29 DF,  p-value: < 2.2e-16
```

Data Analytics

51

Here we have removed these four samples 4, 13, 35 and thing and run the regression analysis again. You can see that the regression analysis maintain retaining all the samples is shown on the right hand side the plot the and their corresponding intercept coupon the slope as well as the F-test statistic and so on R squared values shown here.

And, once we remove these four samples which we outliers and then rerun it now the fit trims to be much much better. It is also seen on the left hand side that the fit is much better. You can see that the R squared value has gone up to 0.99 from 0.75 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 take null hypothesis that a reduce model is adequate which means the linear model with the slope parameter is a much better fit. So, all of these indicators so seems to show that a linear model is adequate and the fit 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 there are no outliers and therefore, we can conclude that the regression model, that were fitted for this data is a reasonably good one.

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

Thank you.