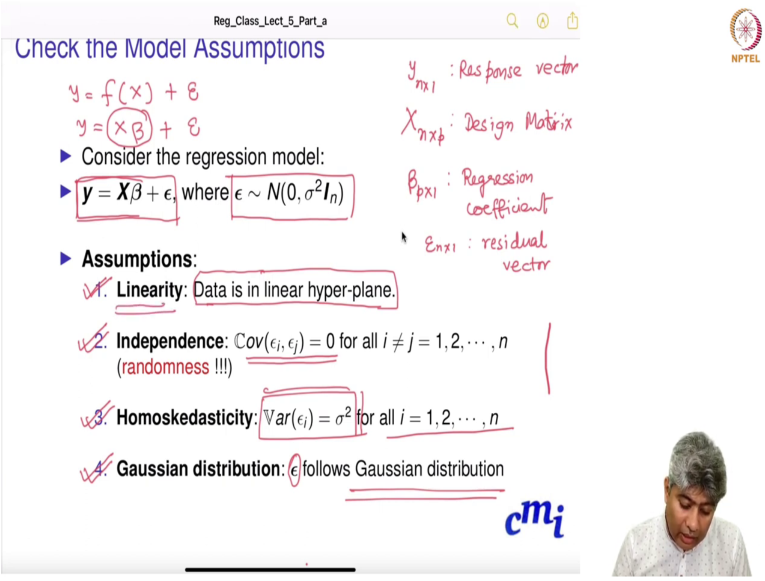
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 17
Checking Model Assumptions

Welcome back to Predictive Analytics-Regression Classification lecture series and today we are going to start Lecture 5 part A.

(Refer Slide Time: 00:26)



Check the Model Assumptions

$y = f(x) + \varepsilon$

$y = \boxed{X\beta} + \varepsilon$

- Consider the regression model:
- $\boxed{\mathbf{y} = \mathbf{X}\beta + \epsilon}$, where $\boxed{\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)}$
- **Assumptions**:
  1. **Linearity**: Data is in linear hyper-plane.
  2. **Independence**: $\mathbb{C}ov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j = 1, 2, \cdots, n$ (randomness !!!)
  3. **Homoskedasticity**: $\boxed{\mathbb{V}ar(\epsilon_i) = \sigma^2}$ for all $i = 1, 2, \cdots, n$
  4. **Gaussian distribution**: ϵ follows Gaussian distribution

$y_{n\times 1}$: Response vector

$X_{n\times p}$: Design Matrix

$\beta_{p\times 1}$: Regression coefficient

$\varepsilon_{n\times 1}$: residual vector

In this lecture we are going to discuss the how to check model assumptions. So, as usual we start with the model like y equal to X beta plus epsilon linear model or linear regression model with epsilon follow normal 0 sigma square I n. So, I am writing it as a I am writing it as a matrix notation and remember that y is n cross 1 vector response vector response vector X is n cross p matrix known as Design matrix.

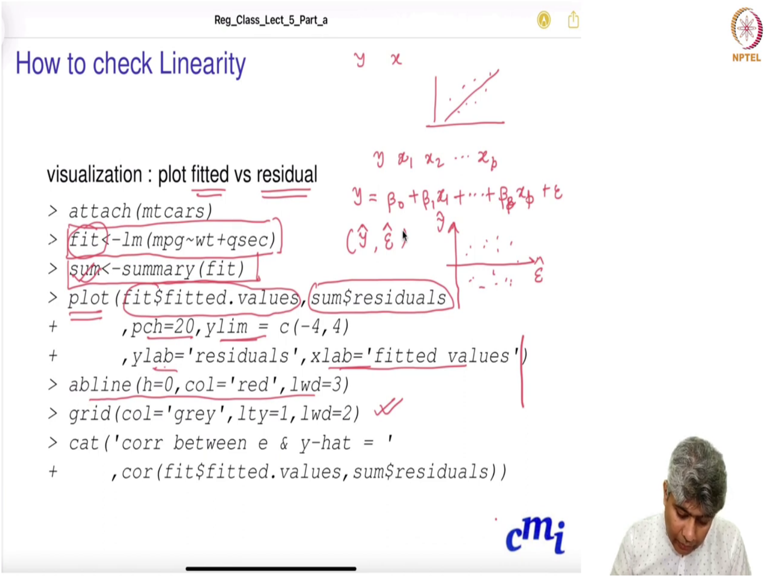
Typically it will be containing the covariates predictors or sometimes it called features. So, this n cross p beta is the regression coefficient or weight matrix typically called Regression coefficient and epsilon is the n cross 1 residual vector residual vector ok, a vector of size n of residual or the part of the model which you cannot explain. That the model has 4 assumptions one is the Linearity, that means data is in a linear hyper plane.

So, this is this you have a you can think of that model is actually some functional form of x plus epsilon this function could be any function, but when you are saying that y is equal to X beta plus epsilon, then you have having a particular form of model and that is called linear models and that is the major assumption. And how do you check whether the linearity that you assuming in the hyper plane, the linear hyper plane that you are assuming for this data is it a right assumption.

Second assumption is Independence and when you assume independence; that means, basically ith observation the residuals of the ith observations and residual of the j th observation they are covariance is 0 or they do not have any correlation. That means, they will not have any correlation they are independent that is a conditionally they may be dependent depend, but the residuals will be independent.

Third assumptions is known as Homoskedasticity; that means, all residual have same variance all residuals have the same variance ok for all i 1 to n and then the fourth assumption is Gaussian distribution. That epsilon this epsilon follow normal 0 sigma square see this Gaussianity is the one of the major fundamental assumption that we are making. So, the first we will discuss how to check a how to check the linearity or data that is data is in hyperplane. So, how to check the linearity?

(Refer Slide Time: 04:32)



How to check Linearity

$y \quad x$

$y \; x_1 \; x_2 \cdots x_p$

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$

$(\hat{y}, \hat{\varepsilon})$

$\hat{y}$

$\hat{\varepsilon}$

visualization : plot fitted vs residual

```
> attach(mtcars)
> fit<-lm(mpg~wt+qsec)
> sum<-summary(fit)
> plot(fit$fitted.values,sum$residuals
+       ,pch=20,ylim = c(-4,4)
+       ,ylab='residuals',xlab='fitted values')
> abline(h=0,col='red',lwd=3)
> grid(col='grey',lty=1,lwd=2)
> cat('corr between e & y-hat = '
+     ,cor(fit$fitted.values,sum$residuals))
```

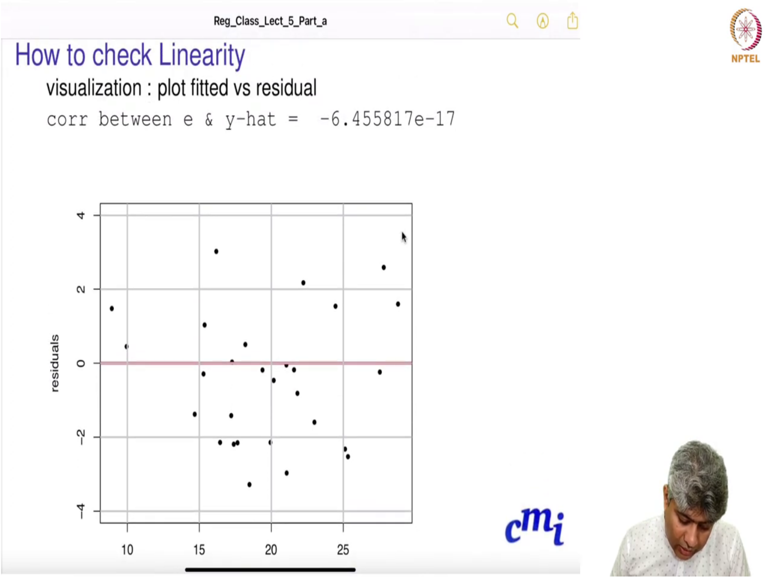
So, it is difficult particularly in maybe if you have only 2 variables say y and x then you can like you simply draw the plot and you can check whether they are in a linear or not linear. But if you have multiple independent variables like you have data with y x 1 x 2 dot dot dot x p, p many variables; that means, your data is in a p plus 1 dimensional geometry then it is very difficult whether the data is the model is in really a linear hyper plane or not.

Now, how what can you do one visualization technique is you plot the fitted values versus the residual values. So, you plot the y you first fit the model beta naught beta 1 x 1 plus dot dot dot beta 2 x beta p x p you fit this model plus epsilon. And the fitted model you calculate the y hat and the epsilon hat and then you plot these 2 plot epsilon hat on the x axis and y hat on the y axis are the and once you if you if you do not have any pattern, then you can say that it is actually probably a good you know that linearity assumption is probably a valid assumptions ok.

So, what we have done that first we fit the model using l m function remember and then we are calling it ok give me the summary of the fit and then what we are doing from the fit we are taking the fitted values and from the summary from the summary we are taking the residual values ok. And then we are calling we are saying that plot these 2 these are the graphical features bunch of graphical features.

We have given y labels x labels the limit of the y axis and all these things and we draw ab line ab line at the horizontal line and we have also asked to draw a grid and we plot some of the things ok.

(Refer Slide Time: 07:31)

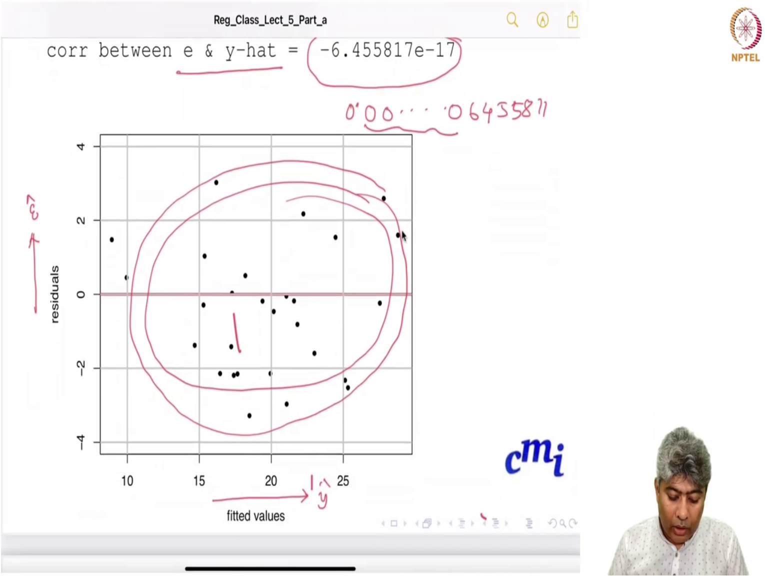


How to check Linearity

visualization : plot fitted vs residual

```
corr between e & y-hat =  -6.455817e-17
```

(Refer Slide Time: 07:43)



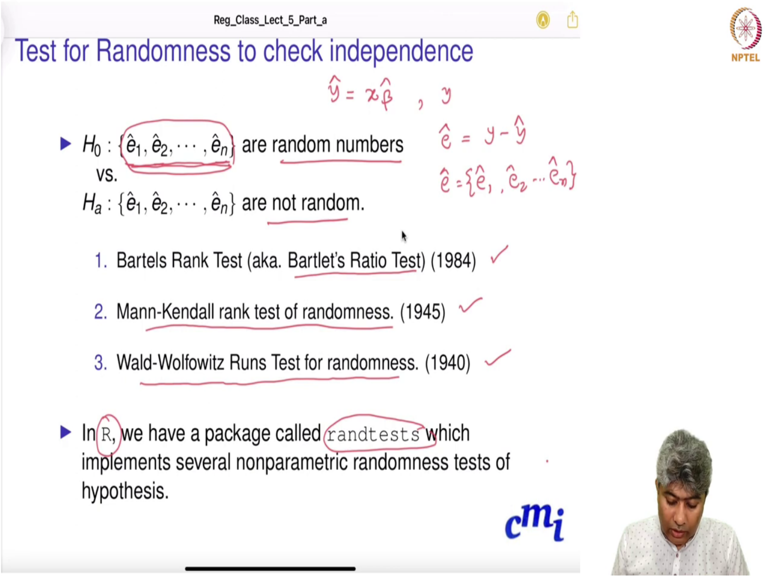
```
corr between e & y-hat = -6.455817e-17
```

0'00··· 06455817

Now when we run this piece of r corr, if we run this piece of r corr what we find is this on the x axis we plot the fitted values the y hats and the y axis we fit the residuals the error. And we can more or less say that you know there is no major correlation or major pattern that we see in this fit and we also calculated the correlation between the you know residual and y hat and we found that they are almost 0.

I mean it is all practical purpose you know that its saying basically after decimal you put 17 0's and then 6 4 5 5 8 1 7. So, all practical purpose if you round it even up to 8 decimal places it is 0. So, there is we cannot find any correlation between the between the residuals and y hat. So, and also visually we see that there is no major pattern between the 2 ok.

(Refer Slide Time: 09:00)



Reg_Class_Lect_5_Part_a

Test for Randomness to check independence

$\hat{y} = x\hat{\beta}$, y

$\hat{e} = y - \hat{y}$

$\hat{e} = \{\hat{e}_1, \hat{e}_2 \cdots \hat{e}_n\}$

- $H_0 : \{\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_n\}$ are random numbers
  vs.
  $H_a : \{\hat{e}_1, \hat{e}_2, \cdots, \hat{e}_n\}$ are not random.

  1. Bartels Rank Test (aka. Bartlet's Ratio Test) (1984) ✓
  2. Mann-Kendall rank test of randomness. (1945) ✓
  3. Wald-Wolfowitz Runs Test for randomness. (1940) ✓

- In R, we have a package called `randtests` which implements several nonparametric randomness tests of hypothesis.

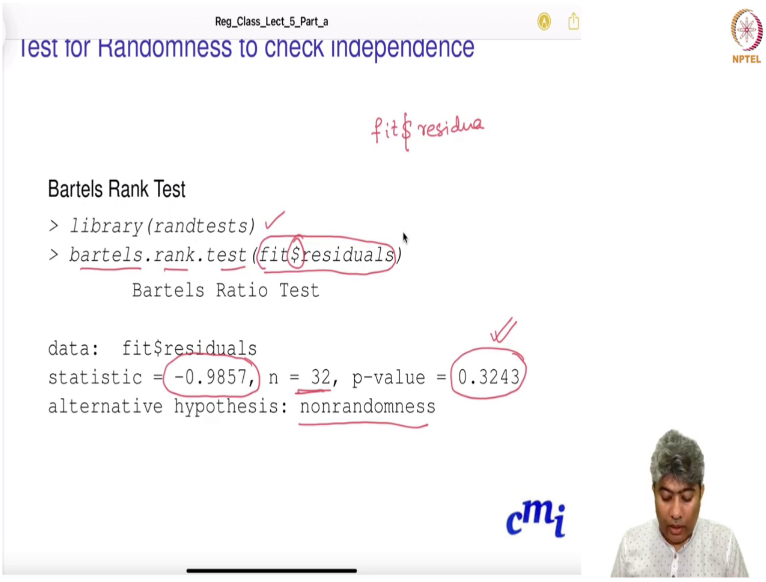
Next we will check the second assumption Test for Randomness to check independence. So, what is our null hypothesis what we will do here? After fitting the model we have these errors right we have this errors e 1 hat e 2 hat e n hat. So, this is after fitting the model we have y hat which is x beta hat and then we also have the y.

So, we can compute the e hat from y minus y hat ok, now e hat is basically our e 1 hat e 2 hat dot dot dot e n hat. Now question is this number are these numbers is this vector the numbers of this vector are random numbers or not random numbers that is the question that we are going to ask. So, the null hypotheses we are saying that yes these are indeed random numbers and alternative hypothesis these are not random numbers.

So, there are 3 popular tests are there one is Bartels Rank Test 1984 Bartels also known as Bartels Ratio Test. Second is Mann-Kendall rank test for randomness and the third is Wald

Wolfowitz Run Test for randomness. So, in which Bartels Rank Ratio test is comparatively newer test with a better statistical power. So, from the R package we have in the R package we have called the rank test package to which implement the several non parametric randomness test for of hypothesis ok.

(Refer Slide Time: 11:19)



Test for Randomness to check independence

fit$ residua

Bartels Rank Test

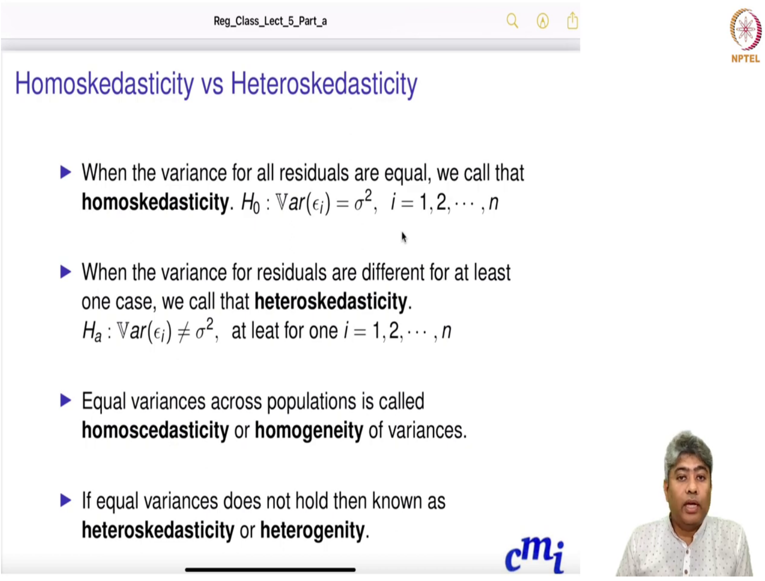
```
> library(randtests)
> bartels.rank.test(fit$residuals)
        Bartels Ratio Test

data:  fit$residuals
statistic = -0.9857, n = 32, p-value = 0.3243
alternative hypothesis: nonrandomness
```

So, we have done here the Bartels Rank Test because it is a new test comparative the more popular ones and we know that statistical power for Bartels rank test is better than the others. So, we called the library rank tests and in that rank test there is a function called Bartels dot rank dot test and you just supply the residual values from the fit.

So, if you basically I am saying we have done this fitted model from that you extract the residuals that is what this dollar sign is basically extract the residuals as a vector and supply this as a vector in the Bartels rank test then it will run the test the statistics is negative 0.9857 sample size is 32 and the p value is 0.3243.

In that in this test alternate hypothesis is non randomness. So, the null hypothesis is randomness and alternate hypothesis is non randomness and p value is since quite large. So, we fail to reject the null hypothesis. So, we stick to in this particular test we stick with the null hypothesis that these residuals are random indeed.

(Refer Slide Time: 12:56)



Homoskedasticity vs Heteroskedasticity

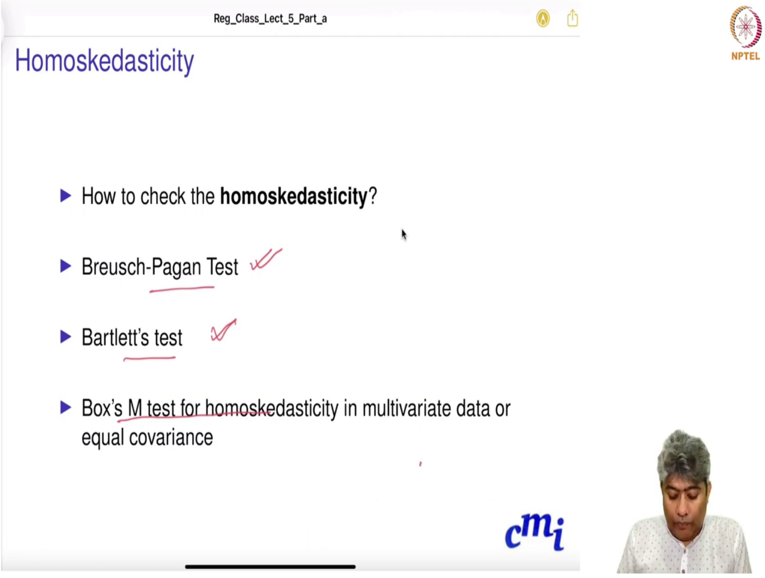
- When the variance for all residuals are equal, we call that **homoskedasticity**. $H_0 : \mathbb{V}ar(\epsilon_i) = \sigma^2, \;\; i = 1, 2, \cdots, n$
- When the variance for residuals are different for at least one case, we call that **heteroskedasticity**. $H_a : \mathbb{V}ar(\epsilon_i) \neq \sigma^2$, at leat for one $i = 1, 2, \cdots, n$
- Equal variances across populations is called **homoscedasticity** or **homogeneity** of variances.
- If equal variances does not hold then known as **heteroskedasticity** or **heterogenity**.

Next we will move on to the discussion for Homoskedasticity versus Heteroskedasticity. So, when the variance for all residuals are equal we call that homoskedasticity. So, this is what we are talking about. So, homoskedasticity is variance for all the residuals are sigma square this is same for every residuals and this assumption is typically called homoskedasticity.

Now when the variance of residuals are different for at least one case we can call this as a heteroskedasticity and the alternative is called variance of epsilon i is not equal to sigma square at least for at least one i ok. So, this is my alternative hypothesis. So, equal variance across a null hypothesis is talking about equal variance across the population is called homoskedasticity or homogeneity of variances, homogeneity of variances.

If the equal variance does not hold then it is known as heteroskedasticity or simply heterogeneity is. If somebody says that there is a heterogeneity in the data basically; that means, they are saying that the variances of residuals are not same.

(Refer Slide Time: 14:36)

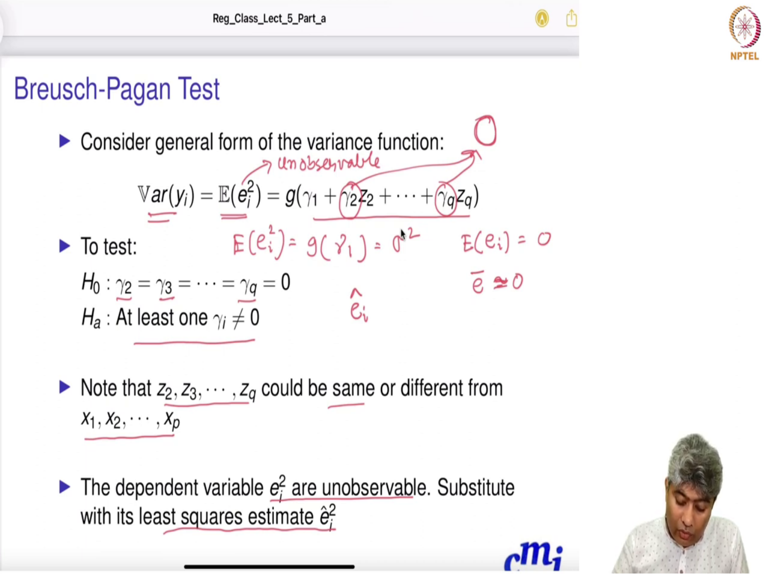


Homoskedasticity

- How to check the **homoskedasticity**?
- Breusch-Pagan Test
- Bartlett's test
- Box's M test for homoskedasticity in multivariate data or equal covariance

So, question is how can we check homoskedasticity? There are quite a few many actually tests are available one very popular test is Breusch-Pagan Test then, Bartlett's test Box's M test for homoskedasticity in multivariate data is also available ok. So, these are multiple test are available.

(Refer Slide Time: 15:09)



Reg_Class_Lect_5_Part_a

Breusch-Pagan Test

- Consider general form of the variance function:

$$\mathbb{V}ar(y_i) = \mathbb{E}(e_i^2) = g(\gamma_1 + \gamma_2 z_2 + \cdots + \gamma_q z_q)$$

unobservable

0

- To test:

$E(e_i^2) = g(\gamma_1) = \sigma^2$ $E(e_i) = 0$

$H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_q = 0$

$\bar{e} \approx 0$

H_a : At least one $\gamma_i \neq 0$

$\hat{e}_i$

- Note that $z_2, z_3, \cdots, z_q$ could be same or different from $x_1, x_2, \cdots, x_p$

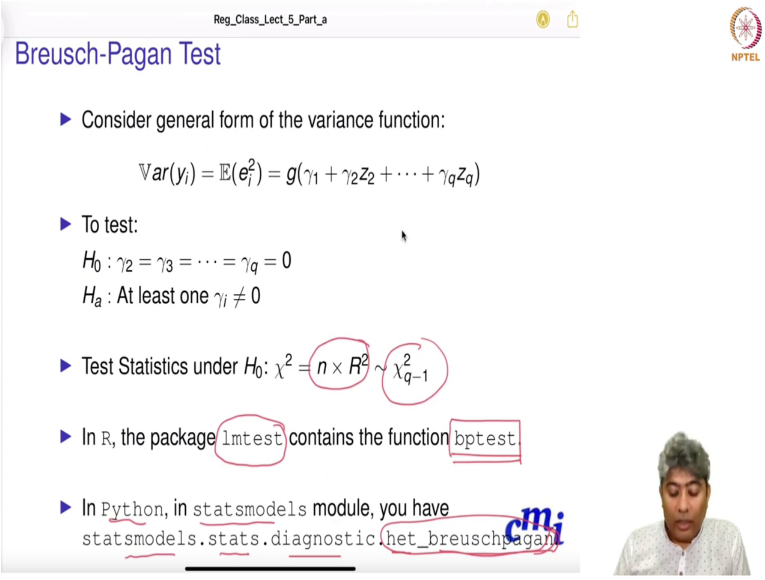
- The dependent variable e_i^2 are unobservable. Substitute with its least squares estimate $\hat{e}_i^2$

Now first we will tell discuss what is Breusch-Pagan Test. So, consider a general form of variance ok and variance of yi is essentially expectation of ei square, because you can effectively say expectation of ei is 0, particularly if you are using OLS method then you can show that e bar the sample mean will be very close to approximately equal to 0 you can pretty much show that.

Now what here we are thinking is to show this e square expected value of residual square is some function of z 1 z 2 z q. What are the z? z 1 z 2 z 3 these are could be the same or different from x 1 x 2 xp or some functional form of x 1 x 2 xp, but some predictors ok. Now, if all the gamma 2 gamma 3 gamma q all these guys are 0, if all of them are 0 then what happens then I am saying basically expected value of e i square is equal to g of gamma 1 a constant right some sigma square ok. But if at least one of gamma is non zero then it is not same.

So, this is the test that we are going to do that gamma 2 gamma 3 gamma q are all 0 and at least one gamma i is not 0. The dependent variable ei square are unobservable remember that, but substitute with the least square estimate ei hat square. So, this is unobservable ok this is unobservable. So, what we do, we instead of directly working with the ei square we work with the ei hat ok. So, that is what we do.

(Refer Slide Time: 17:37)



Breusch-Pagan Test

- Consider general form of the variance function:

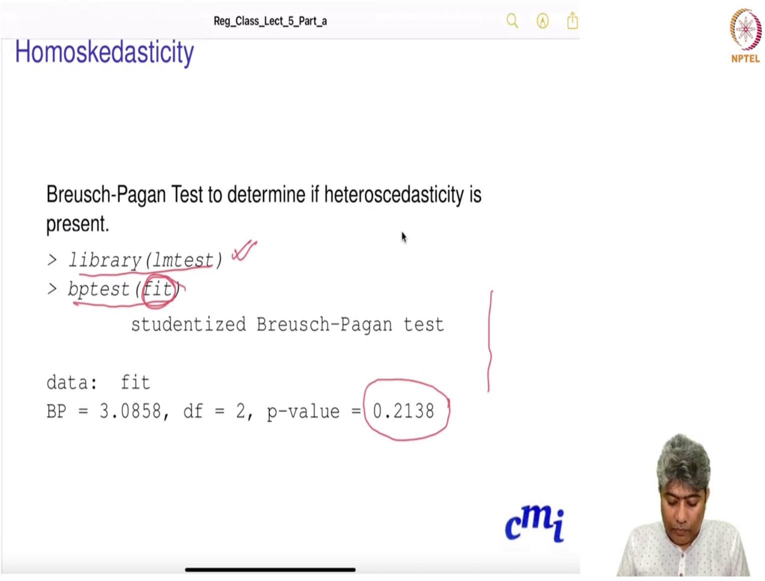
$$\mathbb{V}ar(y_i) = \mathbb{E}(e_i^2) = g(\gamma_1 + \gamma_2 z_2 + \cdots + \gamma_q z_q)$$

- To test:
  $H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_q = 0$
  H_a : At least one $\gamma_i \neq 0$
- Test Statistics under H_0: $\chi^2 = n \times R^2 \sim \chi^2_{q-1}$
- In `R`, the package `lmtest` contains the function `bptest`
- In `Python`, in `statsmodels` module, you have `statsmodels.stats.diagnostic.het_breuschpagan`

Now what turns out that the R square that we there is a chi square statistics that n times R square it follows chi square distribution with q minus 1 degrees of freedom and in the lmtest package in R it contains a function called bptest stands for Breusch-Pagan Test. In python in stat models module if you have the you have to call from the stats model you have to call the stats and then from diagnostics you have to call this het, het stands for heterogeneity Breusch-Pagan.

So, test for heterogeneity using Breusch-Pagan. So, in python also the Breusch-Pagan test do available. So, either you can use i mean python or you can use R either way you can run the Breusch-Pagan test it should give you the same thing.

(Refer Slide Time: 18:50)



Homoskedasticity

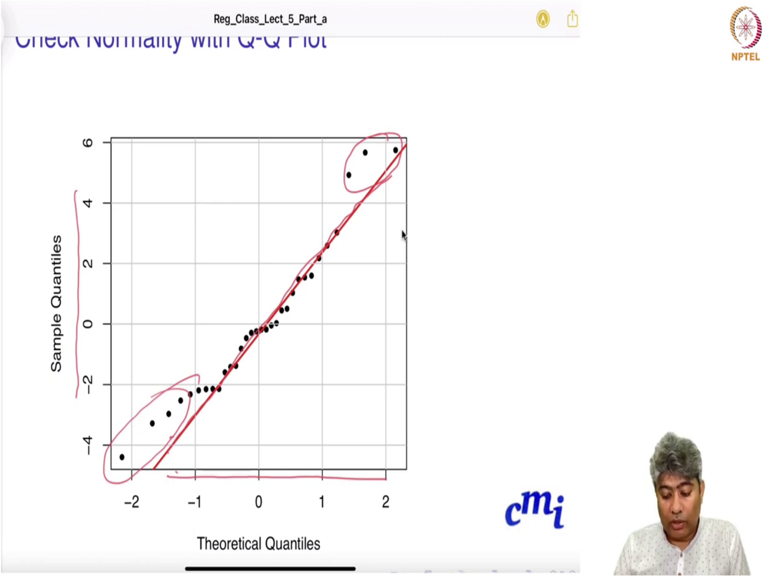
Breusch-Pagan Test to determine if heteroscedasticity is present.

```
> library(lmtest)
> bptest(fit)

        studentized Breusch-Pagan test

data:  fit
BP = 3.0858, df = 2, p-value = 0.2138
```

Now when we run the Breusch-Pagan Test using lmtest in the R, if the heteroscedasticity is present to check that so all we have done its a 2 line code is basically called the library lmtest and run the bptest and then in that you just pass the fit and that will give you the Breusch-Pagan Test. And that is p and what we found that the p value is 2138; 0.2138 which is large enough for which is large enough and hence we fail to reject the null hypothesis.

So, we since we fail to reject the null hypothesis, so we can safely say that at least for this test this model the Breusch-Pagan Test is fail to reject the homoskedasticity hence assumption of homoskedasticity is reasonable for this particular data set for this particular model.

(Refer Slide Time: 20:04)

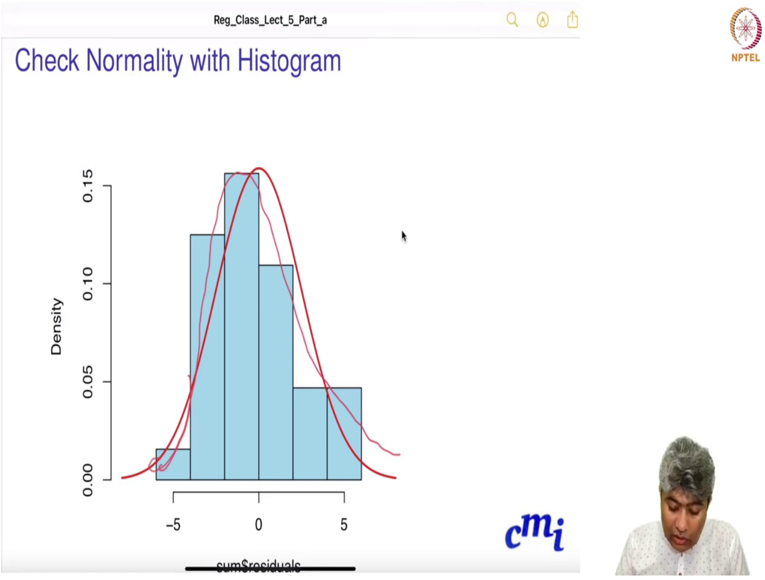


Check Normality with Q-Q Plot

Finally we will check the fourth assumption called check normality for Q-Q Plot, Now Q-Q Plot is very popular all you have to do that calculate the quantiles theoretical quantiles of normal distribution and you calculate the sample quantiles from the data plot them and if they are indeed coming from the same distribution, then they should be on the same line.

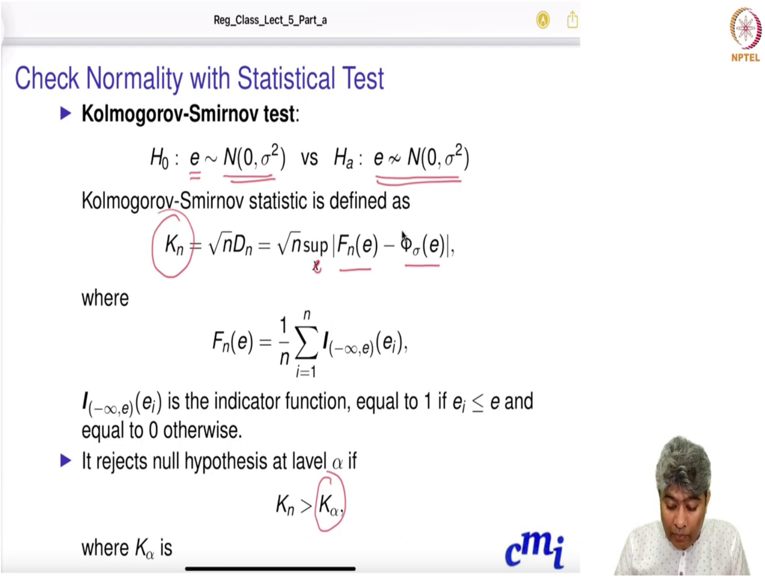
So, that is what you do. So, and in this case visually looks like they are on the same line; however, on the tail side these points are looks like going far away from the line. So, looks like normality could be a questionable thing.

(Refer Slide Time: 21:02)



So, let us try to check if how the residuals histogram of the residuals looks like. Now what we are seeing since we have only 32 observations it looks like that it is little bit maybe off from the normality, but it is bit difficult to say.

(Refer Slide Time: 21:31)



Check Normality with Statistical Test

- **Kolmogorov-Smirnov test**:

$$H_0 : \underline{\underline{e}} \sim \underline{N(0, \sigma^2)} \quad \text{vs} \quad H_a : \underline{e \nsim N(0, \sigma^2)}$$

Kolmogorov-Smirnov statistic is defined as

$$\boxed{K_n} = \sqrt{n}D_n = \sqrt{n}\underset{e}{\sup}|\underline{F_n(e)} - \underline{\Phi_\sigma(e)}|,$$

where

$$F_n(e) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(-\infty, e)}(e_i),$$

$\mathbf{I}_{(-\infty, e)}(e_i)$ is the indicator function, equal to 1 if $e_i \leq e$ and equal to 0 otherwise.

- It rejects null hypothesis at lavel α if

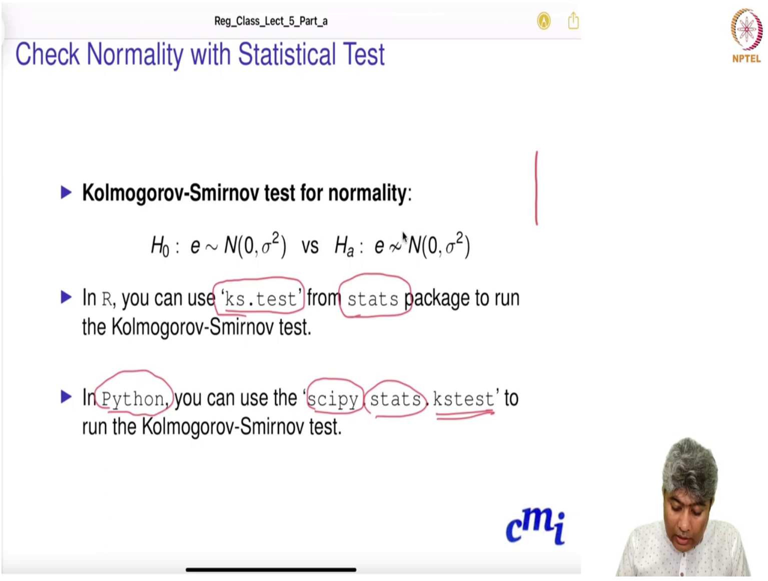
$$K_n > \boxed{K_\alpha},$$

where K_α is

So, what we can we do a test of hypothesis to check the normality. So, to check the normality with statistical test the most popular one perhaps the Kolmogorov Smirnov test. What basically assumes that the residual follow normal 0 sigma square versus residual does not follow normal 0 sigma square. So, the Kolmogorov Smirnov test statistics is essentially square root of n and supremum of that and that will give you the Kolmogorovs Smirnov test statistics.

And if it there will be a k alpha level critical point and if the Kn is greater than K alpha you reject the null hypothesis. So, that is the way we do Kolmogorovs Smirnov test at a alpha level percent.

(Refer Slide Time: 22:32)



Check Normality with Statistical Test

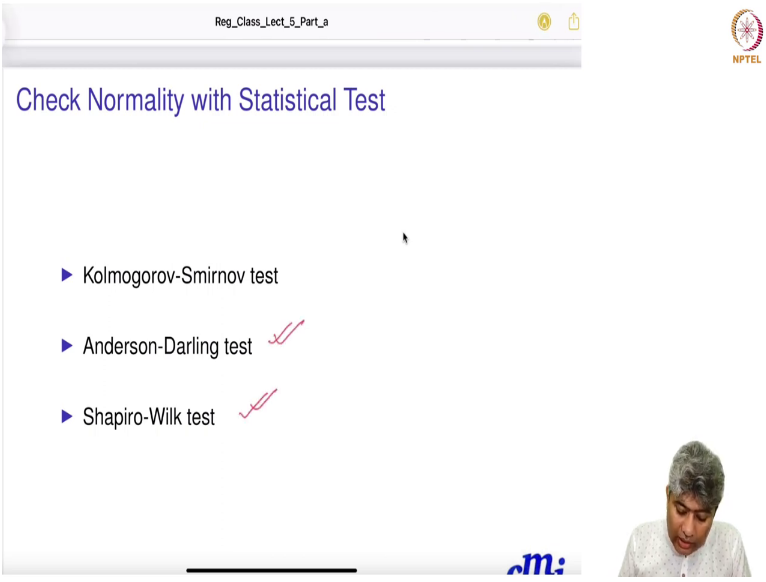
- **Kolmogorov-Smirnov test for normality**:

$$H_0: e \sim N(0,\sigma^2) \quad \text{vs} \quad H_a: e \nsim N(0,\sigma^2)$$

- In `R`, you can use '`ks.test`' from `stats` package to run the Kolmogorov-Smirnov test.
- In `Python`, you can use the '`scipy.stats.kstest`' to run the Kolmogorov-Smirnov test.

So, in r in R you have ks dot test you have ks dot test from the stats package and just run the Kolmogorovs Smirnov test. And in python from scipy you just call the stats module and from there you just have the ks test to run the Kolmogorov Smirnov test. So, in both r and python you have all these.

(Refer Slide Time: 23:07)

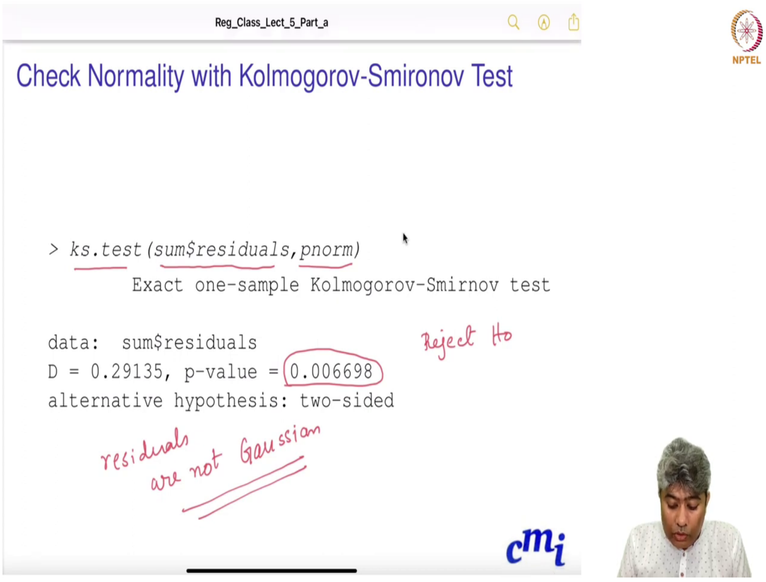


Check Normality with Statistical Test

- Kolmogorov-Smirnov test
- Anderson-Darling test ✓
- Shapiro-Wilk test ✓

And so other than Kolmogorov Smirnov test there is also Anderson Darling test, Shapiro-Wilk test there are also other test tests are available.

(Refer Slide Time: 23:19)



Check Normality with Kolmogorov-Smironov Test

```
> ks.test(sum$residuals,pnorm)
        Exact one-sample Kolmogorov-Smirnov test

data:  sum$residuals
D = 0.29135, p-value = 0.006698
alternative hypothesis: two-sided
```

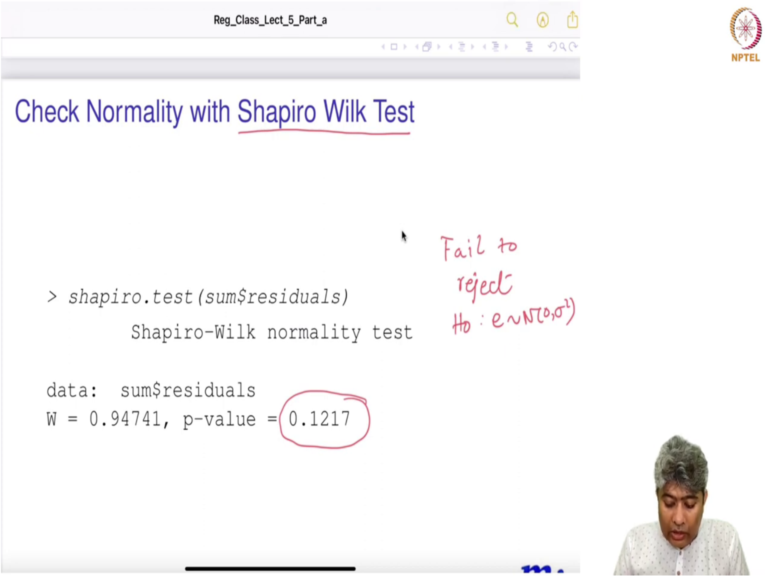
Reject Ho

residuals are not Gaussian

So, I did run the Kolmogorov Smirnov test for residuals and that is basically ks dot test I just subtracted the residuals and then run give that pnorm basically I am asking Kolmogorov Smirnov test to run it against the normality and exact one sample Kolmogorov Smirnov test was run. And what we found that p value is 0.006698 this is quite smaller number. So, we have to reject the null hypothesis reject null hypothesis that is residuals are not normal residuals are not Gaussian distribution not Gaussian.

So, this is this is and you often find that you know we have I have often find in my you know career that many times residuals actually does not call a Gaussian. So, for example, financial data it rarely follow I have seen in financial data rarely follow Gaussian distribution.

(Refer Slide Time: 23:49)



Check Normality with Shapiro Wilk Test

```
> shapiro.test(sum$residuals)

        Shapiro-Wilk normality test

data:  sum$residuals
W = 0.94741, p-value = 0.1217
```

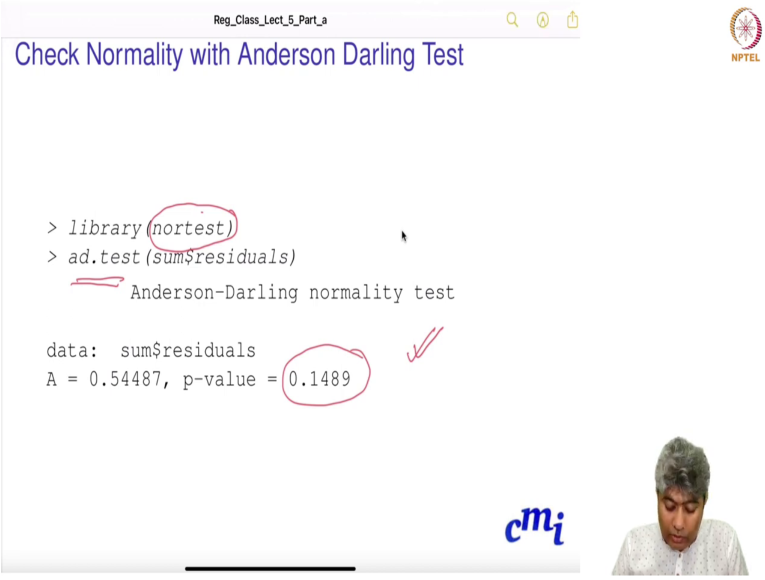
Fail to reject

$H_0 : e \sim N(0,\sigma^2)$

So, it is a very strong assumption in simply regular classical regression methods that residuals follow a Gaussian distribution. Now, there is an alternative method alternative test appears that is called Shapiro Wilk Test, when we run the Shapiro Wilk test we found that p value is 0.1217. So that means, Shapiro Wilk test fail to reject fail to reject null hypothesis ok a null hypothesis.

So, that is that means it is saying that you can probably assume that they are normal it is Shapiro Wilk test is bit conservative in that sense in terms of rejecting null.

(Refer Slide Time: 25:50)

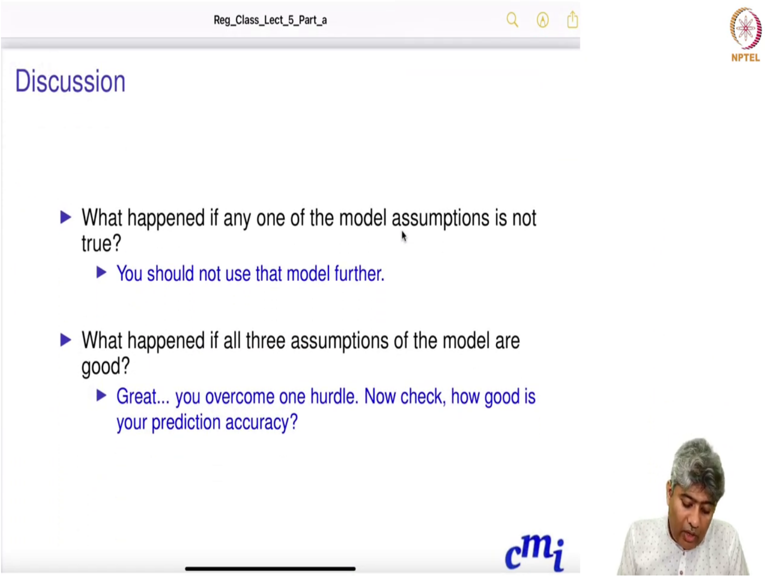


Check Normality with Anderson Darling Test

```
> library(nortest)
> ad.test(sum$residuals)
            Anderson-Darling normality test

data:  sum$residuals
A = 0.54487, p-value = 0.1489
```

So, there is a doubt definitely Anderson darling test also we have done Anderson darling test also in nor test library, you have ad dot test which is Ander stands for Anderson darling test for normality and here also p value is 0.1489 conservative. So, it fails to reject the null. So, in this case also that it says that residuals are probably following normal distribution we do not have enough evidence to reject the null hypothesis.

(Refer Slide Time: 26:27)



Discussion

- What happened if any one of the model assumptions is not true?
  - You should not use that model further.
- What happened if all three assumptions of the model are good?
  - Great... you overcome one hurdle. Now check, how good is your prediction accuracy?

Now, what happened if any one of the model assumption is not true? If any one of the model assumptions is not true then you should not use that model any further. In fact, my recommendation is you should consider using more advanced model or the model which stands with it is own assumptions. And if we in this particular case that I showed as an example we see that though linearity make sense like probably is ok, assumption of linearity is ok though we found homoscedasticity is ok randomness is ok.

But assumption of Gaussianness or Gussianity could be in question out of 3 tests 2 tests failed to reject, but one test did reject in this case I would say that if you want to be conservative it is better you should consider a slightly alternate method may be a bootstrap statistics method or a Bayesian statistics method to overcome this issue. In later some of the lectures I will talk about how to do bootstrap statistics with linear regression or Bayesian statistics with linear regression.

A next big question what happened if all 3 assumptions of the model are good. So, I will say that is a great news. In fact, I would say congratulation that you have overcome come one hurdle, now your duty is to go and check how good your prediction accuracy. You can fit a model you can fit a model and you found that ok all model assumptions are correct.

But you found that your predictive accuracy only 15 percent are you going to use that model no probably not, that is not a good idea to use that model any further ok. Because yes you have fitted a model with all assumptions correctly you know passed through the test of accurate I mean test of hypothesis model assumptions are correct.

But if the predictive accuracy is completely you know through the floor, if the model predictive accuracy is very bad then obviously that model is has no predictive accuracy and it defeats the entire purpose of developing a predictive model. So, our goal is to develop a predictive model which will stand with it is model assumptions from the data as well as it will have a very high good predictive accuracy.

So, with that I will stop here and we will move to the next part and in the next part we will compare the performance of 2 models. So, we will discuss how we can compare the performance of 2 models and different choices of model selection criteria.