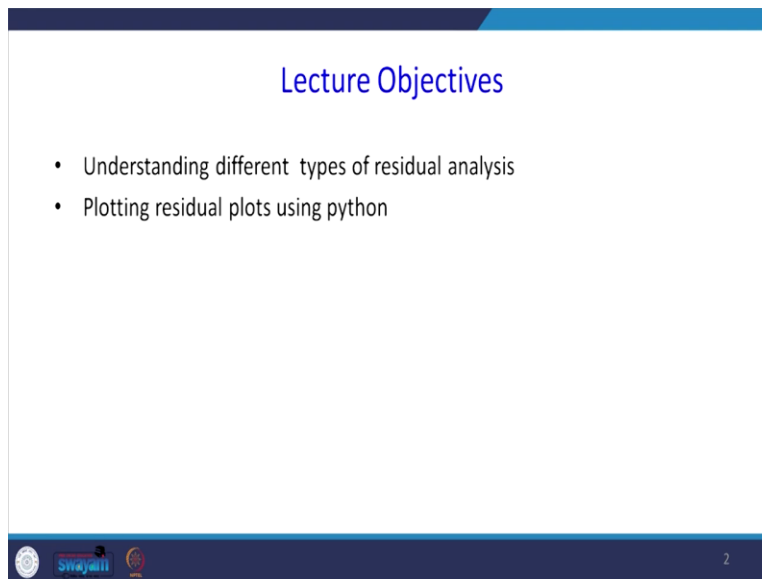


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 32
Estimation Prediction of Regression Model Residual Analysis: Validating Model Assumptions - II

(Refer Slide Time: 00:30)



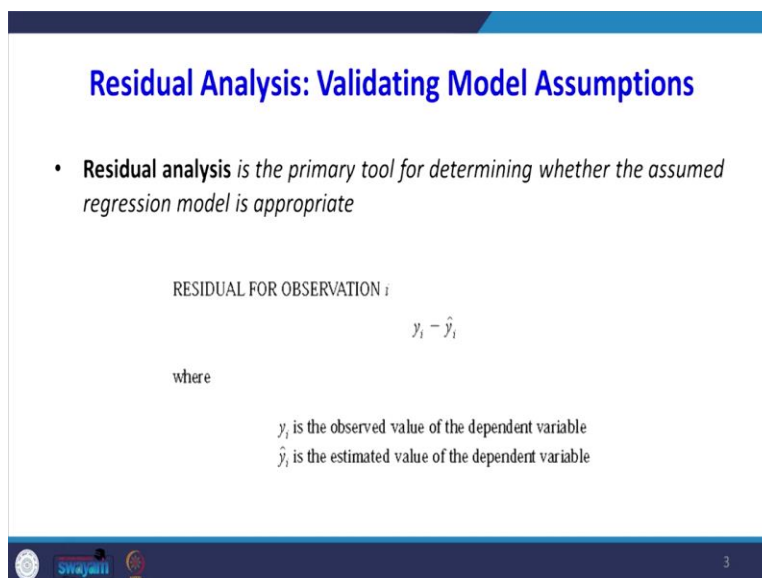
Lecture Objectives

- Understanding different types of residual analysis
- Plotting residual plots using python

2

This lecture is validating regression model assumptions. The lecture objective is understanding different types of residual analysis and plotting residual plots using Python.

(Refer Slide Time: 00:37)



Residual Analysis: Validating Model Assumptions

- **Residual analysis** is the primary tool for determining whether the assumed regression model is appropriate

RESIDUAL FOR OBSERVATION i

$$y_i - \hat{y}_i$$

where

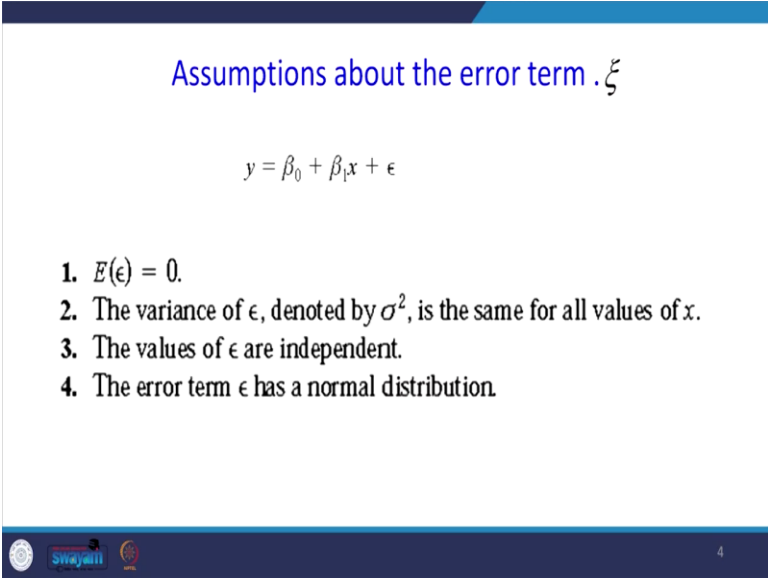
y_i is the observed value of the dependent variable
 \hat{y}_i is the estimated value of the dependent variable

3

Residual analysis validating model assumptions, first we will see what is the residual analysis. The residual analysis is the primary tool for determining whether the assumed regression model is appropriate. So, the residual for observation i is nothing but y_i there is an actual value and \hat{y}_i our predicted model. So, the difference between actual and predicted model it's nothing but the error otherwise you can call it is residual analysis.

So y_i is the observed value of dependent variable \hat{y}_i is the estimated value of dependent variable.

(Refer Slide Time: 01:12)



Assumptions about the error term, ξ

$$y = \beta_0 + \beta_1 x + \epsilon$$

1. $E(\epsilon) = 0$.
2. The variance of ϵ , denoted by σ^2 , is the same for all values of x .
3. The values of ϵ are independent.
4. The error term ϵ has a normal distribution.

4

Assumptions about the error term that is Epsilon we know that y equal to $\beta_0 + \beta_1 x +$ the error term, what are the assumption about this error term, number one the expected value of error is 0. The variance of error term denoted by Sigma square is the same for all values of x the values of error are independent the error term epsilon has their normal distribution. We will validate we will check this assumptions by drawing various residual plots in this lecture.

(Refer Slide Time: 01:47)

Importance of the Assumptions

- These assumptions provide the theoretical basis for the t test and the F test used to determine whether the relationship between x and y is significant, and for the confidence and prediction interval estimates
- If the assumptions about the error term ξ appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

Why this assumption is important these assumptions provide the theoretical basis for the t -test and F test used to determine whether the relationship between x and y is significant and for the confidence interval and prediction interval estimate. What we have done in the previous class we have done t test and F test to test these hypotheses what was our hypothesis: $\beta_1 = 0$ β_1 is $\neq 0$. So, this assumption can be tested by two method only is the t -test and F test so to validate that assumptions the error about assumption is more important.

If the assumptions about the error term ϵ appear questionable the hypothesis test about the significance of the regression relationship and the interval estimation result may not be valid that is why we have to verify this assumptions.

(Refer Slide Time: 02:47)

Residuals for Ice cream parlours

Student Population x_i	Sales y_i	Estimated Sales $\hat{y}_i = 60 + 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	-12
6	105	90	15
8	88	100	-12
8	118	100	18
12	117	120	-3
16	137	140	-3
20	157	160	-3
20	169	160	9
22	149	170	-21
26	202	190	12

Source: Statistics for Business & Economics, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, Cengage Learning, 2013

We will take an example this example is adopted from statistics for Business and Economics David and Anderson Sweeney and Williams the student population x_i 2 6 8 8 12 and so on, sales of ice cream is given so we have fitted the regression line $\hat{y}_i = 60 + 5x_i$ so when you substitute the x value here this is actual 58 this is our predicted 70 the difference is $50 - 70$ is -12 , so 105, 90 the difference is 15, so this $y_i - \hat{y}_i$ is that is the residual. So this residual I have to have some properties that properties we will check it.

(Refer Slide Time: 03:41)

Residual analysis is based on an examination of graphical plots

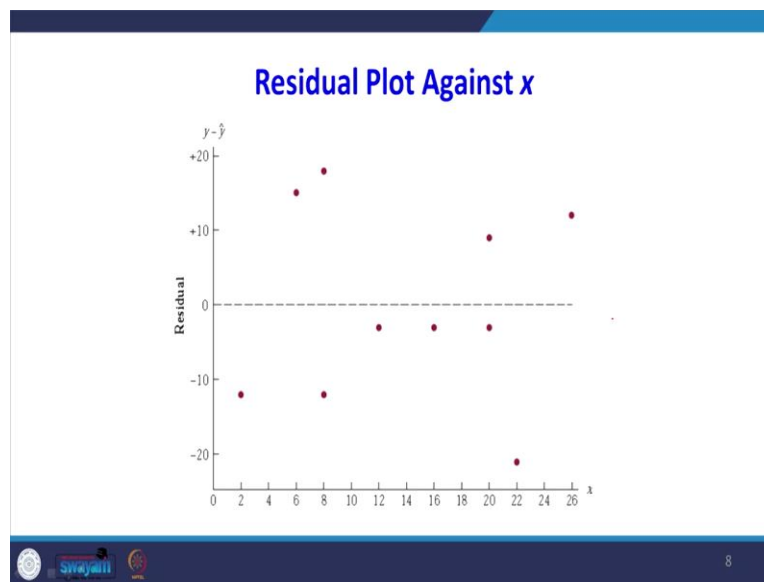
- A plot of the residuals against values of the independent variable x
- A plot of residuals against the predicted values of the dependent variable \hat{y}
- A standardized residual plot
- A normal probability plot

The residual analysis is based on the examination of graphical plot. So, we are going to plot the residual it was in the previous slide then we have to check certain assumptions there are 4 method we are going to do in this class one is a plot of the residual against value of independent

variable x . So, first assumption is x axis we are going to take x value in dependent variable in y axis we are going to take residual the second assumption is the plot of residuals against a predicted value of the dependent variable \hat{y} , so in x axis we are going to have \hat{y} then y axis we are going to have residuals.

The third assumption is standardized as a residual plot we are going to standardize this the residual we know that how to standardize standardized for example if you are standardizing this is $(x - \bar{x})$ by Sigma this is the way Sigma our standard error so it is nothing but z is nothing but $z \cdot t$ so we will standardize our residual then we plot it the last assumption is normal probability plot.

(Refer Slide Time: 04:51)



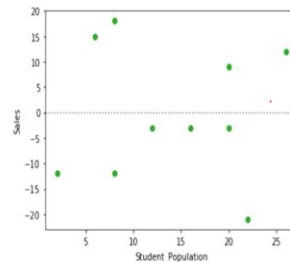
So, we will check this assumptions the first assumption is when we plot the residuals against x value so this error that is duals are plotted in this way. So, what is the inference we can get it number one that it is not following any pattern, if it is not following any pattern these errors are independent then there is no problem in the assumptions.

(Refer Slide Time: 05:15)

Residual Plot Against x

```
In [18]: import seaborn as sns
sns.residplot(df['Student_Population'], df['Sales'], color='g')

Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x24e594e9f60>
```

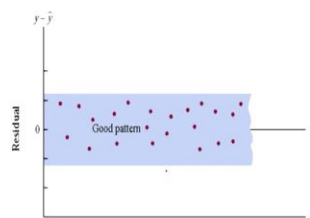


So, this one we have done with the help of Python I have the screenshot at the end of this lecture I am going to run all these codes you can verify it. Import seaborn as sns before that we have to import the data set that I will show you in the end of the class. So, sns.residplot so this is used for plotting that is do one variable a student population that is x value the y value sales color is green color so you will get this output.

So this was the Python output of a residual sales that is a y predicted value against sales y axis not the sales it is the residual, it is the residual, y axis is not the sales because sales would not become 0.

(Refer Slide Time: 06:20)

Assumption: the variance is the same for all values of x

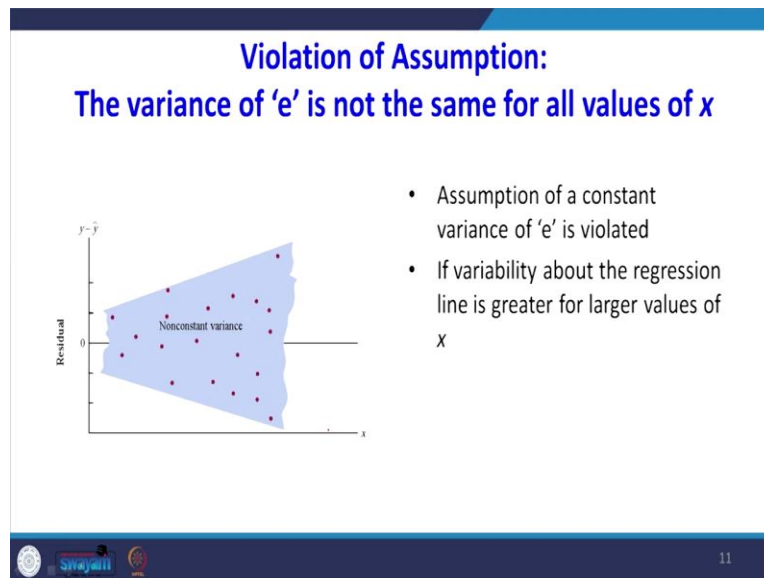


- The residual plot should give an overall impression of a horizontal band of points

So, the one assumption is the variance is the same for all values of x , so what will happen now in this figure which is given it is looking like a rectangle shape that means this assumption is valid it is a good pattern. So, what did this graph implies the residual plot should give an overall impression of horizontal band of points. So, that means the variance even though the x value is increasing the variance is same so then we get here a horizontal band of points so this is the way to check the one assumption that the variance is same for all values of x .

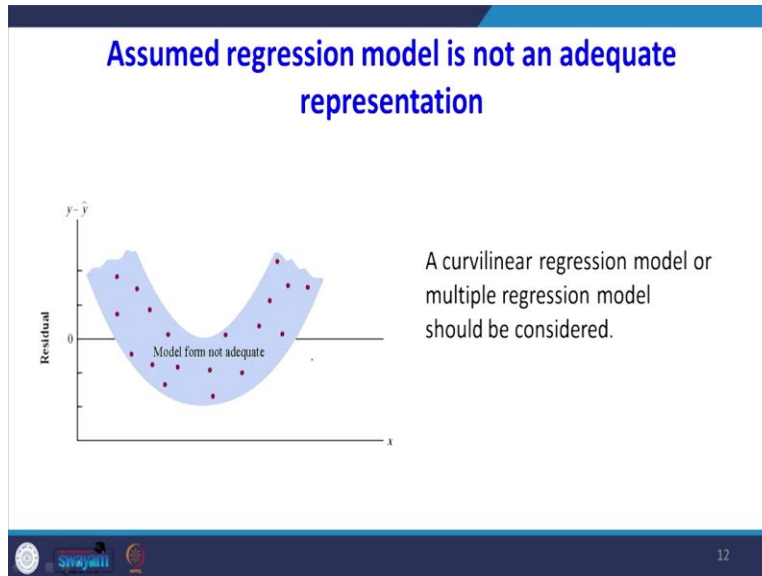
Sometime what may happen when the value of x increases the variance may increase that should not be the case yeah that is an example of this one.

(Refer Slide Time: 06:54)



What is happening violation of assumption what is that the variance of y_i is not the same for all values of x when x is increases the variance is no it is getting a conical shape, so it is a non constant variance. This is the violation of our regression model. So, assumption of a constant variance of E is violated if you are getting this kind of shape. If variability about the regression line is greater for larger values of x then you can get this kind of pictures. So that is not correct one.

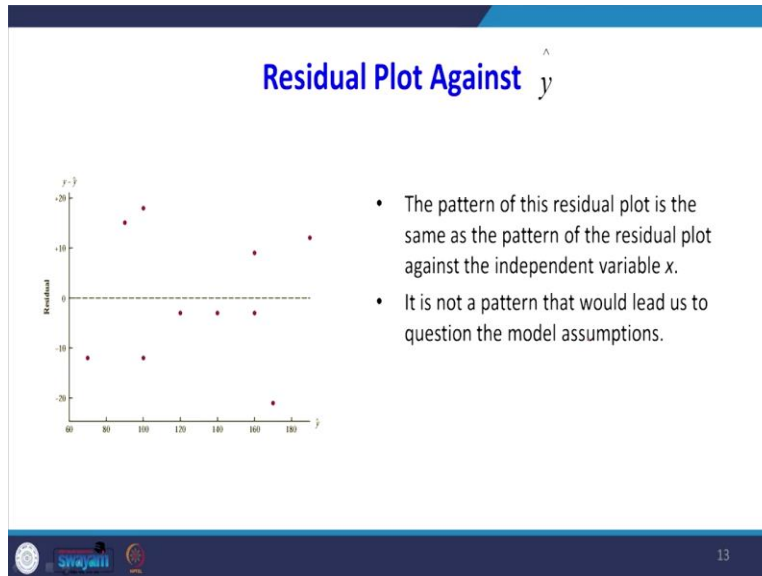
(Refer Slide Time: 07:30)



Another type of picture you may get it when you plot the residual against the x , a curvilinear this is a this is a kind of a non linear shape, so instead of fitting a linear regression equation it is suggesting that you can try for curvilinear regression model or a multiple regression model should be considered if you are getting the plot is in this shape. Previously we have plot the residual against x now we are going to plot the residual against \hat{y} that is our predicted value. The pattern of this residual plot is the same as the pattern of residual plot against an independent variable x .

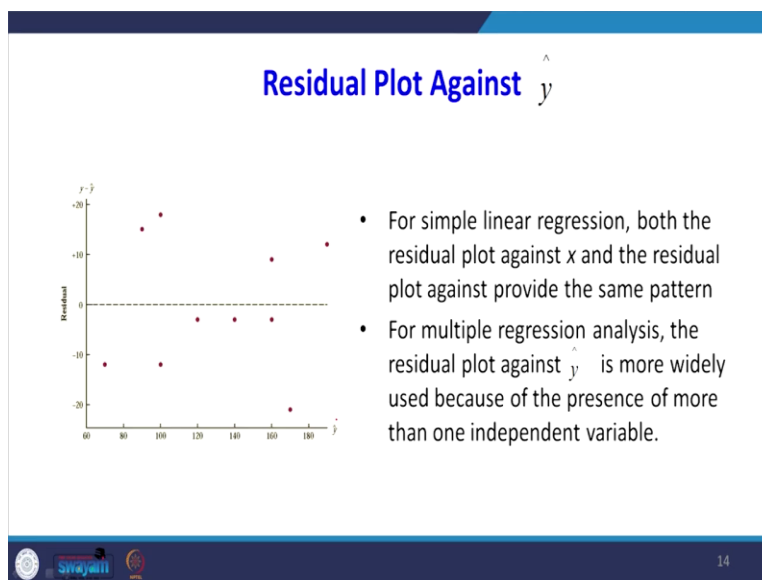
It is not a pattern that that would lead us to question the model assumptions why this we are going for that one if there are more number of independent variable for each independent variable you have plot this residual.

(Refer Slide Time: 08:19)



So, instead of going for different independent variable if you plot this residual against this predicted value then from that we can verify whether the model is valid or not.

(Refer Slide Time: 08:35)



So, for a simple linear regression both the residual plot against x and the residual plot against \hat{y} have provided the same pattern for multiple regression analysis the residual plot against \hat{y} is the more widely used because of the presence of more than one independent variable. Whenever there is more than one independent variable instead of going for x we should go for \hat{y} .

(Refer Slide Time: 08:58)

Standardized Residuals

- Many of the residual plots provided by computer software packages use a standardized version of the residuals.
- A random variable is standardized by subtracting its mean and dividing the result by its standard deviation.
- With the least squares method, the mean of the residuals is zero.
- Thus, simply dividing each residual by its standard deviation provides the **standardized residual**

Then we will go for next residual plot that is a standardized residuals. Many of the residual plots provided by computer software packages uses a standardized version of the residuals. So, what is the standardized version yeah random variable is standardized by subtracting its mean dividing the result by its standard deviation, this way $Z = \frac{\text{residual } i - \text{residual mean value}}{\text{standard deviation of the residual}}$.

With the least square method the mean the residual is 0 because $\sum (x - \bar{x}) = 0$, thus simply dividing each residual by its standard deviation provides the standardized to residual. So, what you have to do in the least square method simply how to divide residual by your standard deviation that will give you the standardized the residual.

(Refer Slide Time: 09:57)

Python Code

```

In [14]: import pandas as pd
         from statsmodels.formula.api import ols
         from statsmodels.stats.anova import anova_lm
         import matplotlib.pyplot as plt

In [9]: df1 = pd.read_excel('Icecream.xlsx')
         df1

Out[9]:
   Student_Population  Sales
0                   2     58
1                   6    105
2                   8     88
3                   8    118
4                  12    117
5                   8    137
6                  20    157
7                  20    169
8                  22    149

In [11]: reg1 = ols(formula = "Sales ~ Student_Population", data = df1)
         fit1 = reg1.fit()
         print(fit1.summary())

```

So, I am so in the Python code here this is a screenshot of the program import pandas as pd, from statsmodels dot formula dot api import OLS from stats model dot stat stat anova import anova underscore lm, import matplotlib dot pyplot as plt. So, the data set name is ice cream so in the independent variable student population the dependent variable is sales. So, to get your regression model reg1 equal to OLS formula equal to sales as a dependent variable tilde student underscore population data equal to df1.

(Refer Slide Time: 10:43)

```

=====
                        OLS Regression Results
=====
Dep. Variable:          y      R-squared:                0.903
Model:                OLS    Adj. R-squared:            0.891
Method:             Least Squares    F-statistic:        74.25
Date:                Thu, 05 Sep 2019    Prob (F-statistic):    2.55e-05
Time:                11:16:42    Log-Likelihood:       -39.342
No. Observations:        10    AIC:                   82.68
Df Residuals:            8    BIC:                   83.29
Df Model:                1
Covariance Type:        nonrobust
=====
               coef    std err          t      P>|t|      [0.025    0.975]
-----
const         60.0000     9.226     6.503     0.000     38.725     81.275
x1             5.0000     0.580     8.617     0.000      3.662      6.338
=====
Omnibus:                 0.928    Durbin-Watson:           3.224
Prob(Omnibus):            0.629    Jarque-Bera (JB):         0.616
Skew:                    -0.060    Prob(JB):                 0.735
Kurtosis:                 1.790    Cond. No.                 33.6
=====

```

So, when you print a summary so you will get this kind of regression output. So, this says your r square this is over adjusted r square I will explain the meaning of our just r square in multiple

regression this was for F statistics. So, what say this is, so y equal to $60 + 5 \times 1 \times 1$ is number of and populations.

(Refer Slide Time: 11:07)

Python Code

```
In [12]: print(anova_lm(Fit1))
```

	df	sum_sq	mean_sq	F	PR(>F)
Student_Population	1.0	14200.0	14200.00	74.248366	0.000025
Residual	8.0	1530.0	191.25	NaN	NaN

18

So, when you use this `print anova_lm` for the our model fit one you can get your ANOVA table for regression analysis, so for you a residual it is 8 because there is a 10 data set so the degrees of freedom is $n - p - 1$, p is number of independent variable there is only one independent variable so the degrees of freedom is 1 this is sum of square for student population this sum of square for error. So sum of squares divided by degrees of freedom you will get mean sum of square.

So the F value is nothing but means sum of squared divided by mean error sum of square so the p value is very low so we can say that the model is valid.

(Refer Slide Time: 11:52)

Standardized Residuals

STANDARD DEVIATION OF THE i th RESIDUAL

$$s_{y_i - \hat{y}_i} = s \sqrt{1 - h_i}$$

where

$s_{y_i - \hat{y}_i}$ = the standard deviation of residual i

s = the standard error of the estimate

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

Next will tell you how to find out these standardized to residual. So, the standardized residual is $y_i - \hat{y}_i$ equal to $s \sqrt{1 - h_i}$ here s is the standard error of the estimate. So, in the previous this is where MSE is 191 when you take the square root of this what is the standard error is standard error is SSE divided by $n - 2$, when you take square root otherwise 1 and 2 and 0.25 when you take square root that you will get the standard error.

So that is nothing but the value of s so you can find out h_i equal to $(1 \text{ by } n) + ((x_i - \bar{x})^2 / \sum (x_i - \bar{x})^2)$.

(Refer Slide Time: 12:41)

Computation of standardized residuals for Icecream parlors

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	-12	144	.2535	.3535	11.1193	-12	-1.0792
2	6	-8	64	.1127	.2127	12.2709	15	1.2224
3	8	-6	36	.0634	.1634	12.6493	-12	-.9487
4	8	-6	36	.0634	.1634	12.6493	18	1.4230
5	12	-2	4	.0070	.1070	13.0682	-3	-.2296
6	16	2	4	.0070	.1070	13.0682	-3	-.2296
7	20	6	36	.0634	.1634	12.6493	-3	-.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	-21	-1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
Total			568					


So, there is an illustration so I use there x_i is there we are finding $x_i - \bar{x}$ because this value will be useful for the formula which is in the previous slide so $(x_i - \bar{x})^2$ whole square so when you know we can $((x_i - \bar{x})^2 / \sum (x_i - \bar{x})^2)$, then you confront a h_i you can find out the $s_{y_i - \hat{y}_i}$ from that you can find out the $y_i - \hat{y}_i$, so then you will get the standardized residual.

(Refer Slide Time: 13:15)

Computation of standardized residuals for Icecream parlors

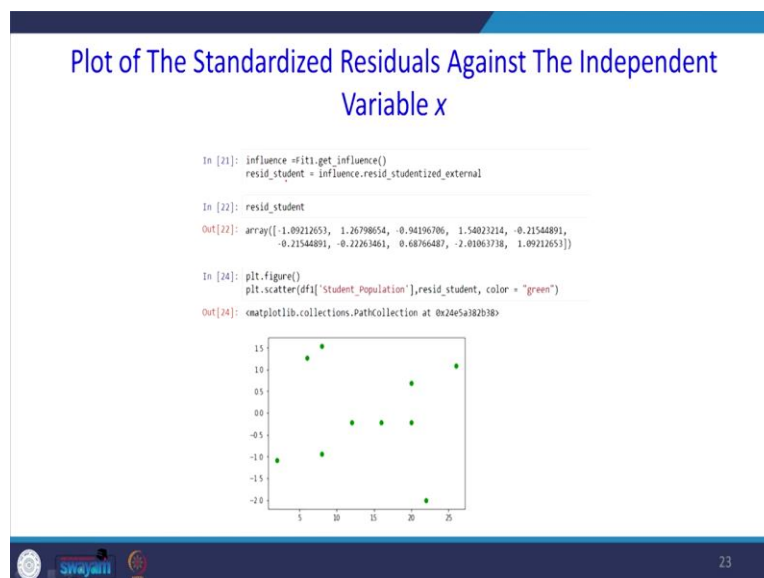
STANDARDIZED RESIDUAL FOR OBSERVATION i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$


21

It is do it so standardized residual is nothing but $y_i - \hat{y}_i / s_{y_i - \hat{y}_i}$ so what will happen when you plot this figure x against this and residual most of the data point is see that between $+2$ and -2 so that means that 95% of the time the data's are within the limit so this is acceptable.

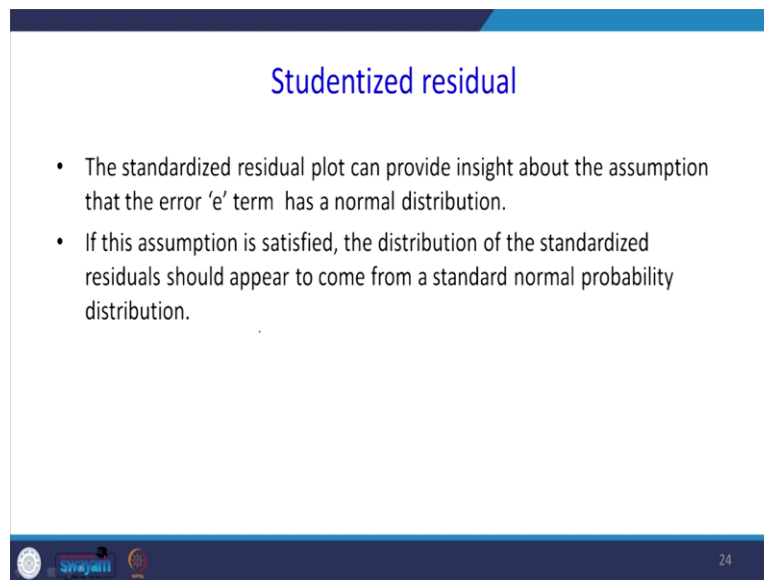
(Refer Slide Time: 13:47)



The assumption is valid we can plot the standardized residual plot against the independent variable x so for that you this command `influence, influence equal to fit1 dot get underscore influence, resid_student equal to influence dot resid_studentized_external, external` so you can see what is that `resid_student` that is an array of this is nothing but the standardized residual.

So now you can plot student population against this studentized residual will get this figure see all the data point is between $+2$ and -2 , so this assumption is valid.

(Refer Slide Time: 14:22)



Studentized residual

- The standardized residual plot can provide insight about the assumption that the error 'e' term has a normal distribution.
- If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.

24

The standardized residual plot can provide insight about the assumption that the error term 'e' has the normal distribution. If this assumption is satisfied the distribution of the standardized residual should appear to come from a standard normal probability distribution.

(Refer Slide Time: 14:41)

Studentized residual

- Thus, when looking at a standardized residual plot, we should expect to see approximately 95% of the standardized residuals between -2 and 2.
- We see in Figure that for the Ice-cream example all standardized residuals are between -2 and 2.
- Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that 'e' has a normal distribution.

Studentized there is dual this when looking at the standardized the residual plot we should expect to see approximately 95% of the standardized residuals between - 2 and + 2. We see the figure that from the ice cream example all standardized residuals are between - 2 and + 2 therefore on the basis of the standardized residuals this plot gives us no reason to question that assumption that the error term has here normal distribution.

(Refer Slide Time: 15:10)

Normal Probability Plot

- Another approach for determining the validity of the assumption that the error term has a normal distribution is the **normal probability plot**.
- To show how a normal probability plot is developed, we introduce the concept of *normal scores*.

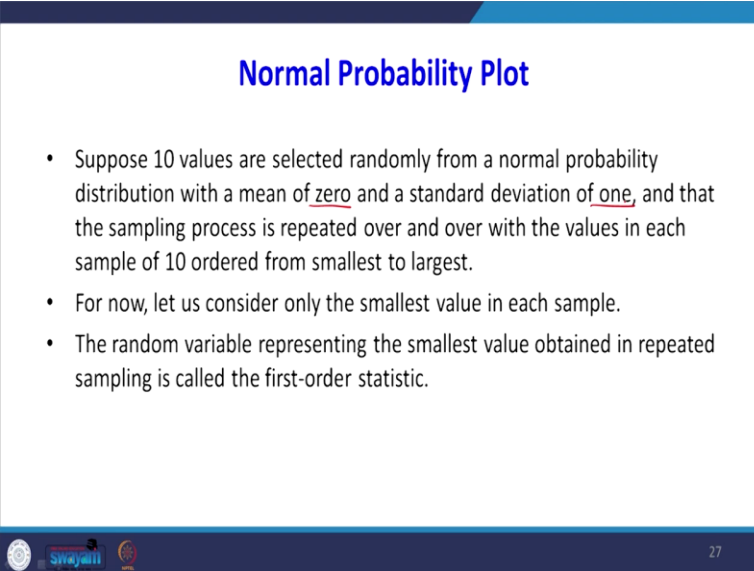
Next we will plot normal probability plot. Another approach for determining the validity of the assumption that the error term has a normal distribution is normal probability plot. Many software packages you may see that the normal probability plot. To show how a normal probability plot is developed we introduce concept called normal score. Suppose 10 values are

selected randomly from a normal probability distribution with the mean 0 and standard deviation 1 and that is sampling process repeated over and over with the values in each sample of 10 ordered from smallest to largest.

(Refer Slide Time: 15:52)

Normal Probability Plot

- Suppose 10 values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of 10 ordered from smallest to largest.
- For now, let us consider only the smallest value in each sample.
- The random variable representing the smallest value obtained in repeated sampling is called the first-order statistic.

The slide features a blue header with the title "Normal Probability Plot". Below the title, there are three bullet points explaining the sampling process and the definition of the first-order statistic. The footer contains a logo on the left and the number "27" on the right.

27

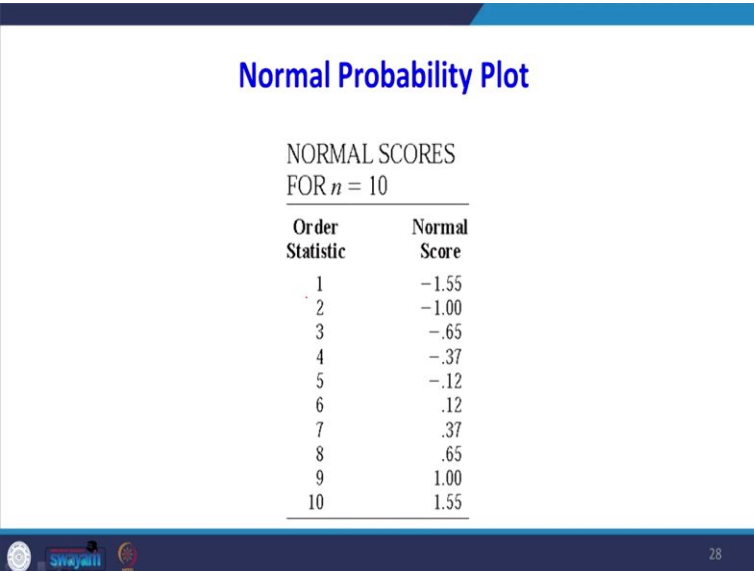
For now let us consider only the smallest value in each sample. The random variable representing the smallest value obtained in a repeated sampling is called first order statistic. Okay the second largest is second order statistic and so on so this was the first order statistics.

(Refer Slide Time: 16:14)

Normal Probability Plot

NORMAL SCORES
FOR $n = 10$

Order Statistic	Normal Score
1	-1.55
2	-1.00
3	-.65
4	-.37
5	-.12
6	.12
7	.37
8	.65
9	1.00
10	1.55

The slide features a blue header with the title "Normal Probability Plot". Below the title, the text "NORMAL SCORES FOR n = 10" is centered. Underneath, there is a table with two columns: "Order Statistic" and "Normal Score". The table lists values for order statistics 1 through 10. The footer contains a logo on the left and the number "28" on the right.

28

So for this first order statistic wherein the sample size equal to 10 it should be - 1.55 that means these values data's are coming from the normal distribution. So, for the second order statistics the

value should be so, this value which we got from the table. Now we are going to compare the standardized residual values with this table so already we have the standardized residual when I equal to 1 x i equal to 2, so it is -1.0792 .

So these values we are going to compare with the standardized so this value we are going to compare with the normal scores.

(Refer Slide Time: 16:57)

Normal Scores	Ordered Standardized Residuals
-1.55	-1.7114
-1.00	-1.0792
-.65	-.9487
-.37	-.2372
-.12	-.2296
.12	-.2296
.37	.7115
.65	1.0792
1.00	1.2224
1.55	1.4230

So, what is happening here then we look at this picture when the order statistic is 1 the minimum value is -1.55 so in this figure we have to see which is near to -1.55 , so the this one -1.71 . So, this is the -1.71 next we have to see which is the least value from this figure next to least value is 1.07 so -1.07 next least. So, we have mapped with this our standardized the residual against the normal score. When it is the the least one is taken as -1.71 when it is 1.55 in the normal score the corresponding score from our dataset is 1.4230 .

(Refer Slide Time: 17:49)

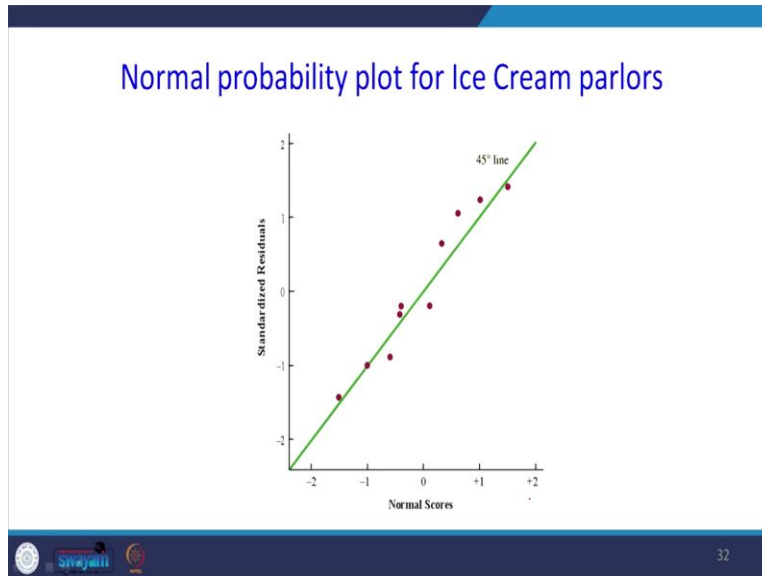
Normal Probability Plot

- If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest normal score, and so on.
- If we were to develop a plot with the normal scores on the horizontal axis and the corresponding standardized residuals on the vertical axis, the plotted points should cluster closely around a 45-degree line passing through the origin if the standardized residuals are approximately normally distributed.
- Such a plot is referred to as a *normal probability plot*.

Now we will see what its normal probability plot by using this data we will plot it if the normality assumption is satisfied the smallest standardized the residual should be close to the smallest normal score. The next smallest standardized residual should be close to the next smallest normal score and so on that is what we mapped it in the previous slides.

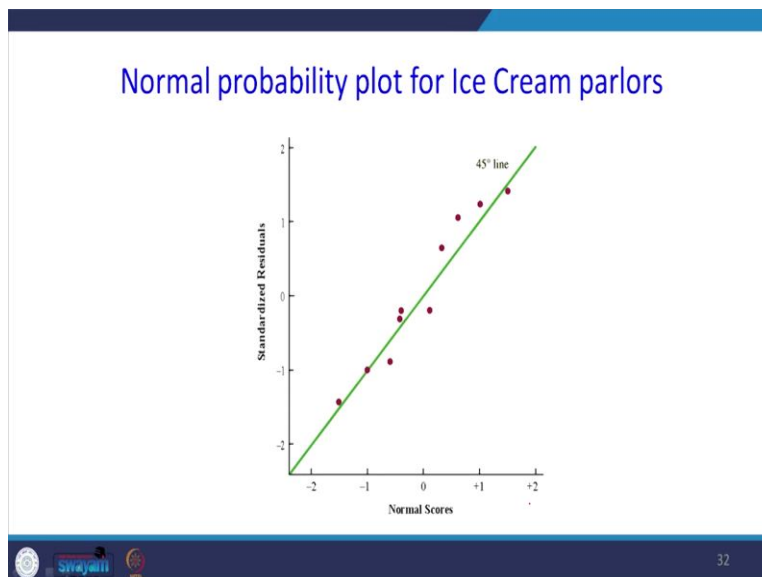
If you had to develop a plot with the normal score on the horizontal axis and the corresponding standardized the residual on the vertical axis the plotted point should cluster closely around the affine is 45-degree passing through the origin. The standardized residuals are approximately normally distributed. It is a property of this residual plot. So, such a plot is referred as normal probability plot.

(Refer Slide Time: 18:36)



So, this was normal probability plot so in x-axis normal score is written in y-axis the standardized residuals written. So, it is starting from 0 the line is 45 degree so all the points are right it is not deviating from this line, so it is it is a clustering around this Green Line. So, then we can say that this data follow normal distribution. Suppose the data is following this by this point this point it is not going it is not clustered around this green line then we can say it is not following normal distribution.

(Refer Slide Time: 19:15)



This also we have done with the help of Python from scipy import stat stats import statsmodels as sm so we are going to take residual as res = fig_1.resid then we go for plot = sm.proplot(res) because sm and all different library which is already important

to see that import statsmodels dot api as sm. So, then figure prompt lot dot qq plot when line equal to 45 degree, so we can say `h= plt.title ('qq plot again to this dual of OLS. fit then we are getting this you see that all the points are above this red line then we can see this normality assumption is validated.`

Now what we are going to do I have prepared this command in Python I go to run all the Python course then I am going to verify I am going to show how to get this residual plots then how to verify that it meet the regression assumptions. So, far I have shown the screenshot of the Python output now I am going to run and I am going to explain how to get the residual plots and what is the interpretation of that.

For that what I have done I have taken one regression example filenames where I have stored the data is ice cream so first I will run this I on the library then I will show what is the data set. This data set shows there is a two variable and this one a student underscore population is independent variable sales is dependent variable. So, for this data set I am going to run the regression equation. We are getting regression output you see that intercept is 60 the intercept of the student underscore population variable is 5.

So we can write it y equal to $60 + 5 \times 1$ here x_1 is student population then we can see that this p-value also it is less than 0.05 so this independent variable is significant values you see that r square is 0.903 that means 90.3% of the variability of Y is explained with the help of this is a regression model. Similarly for the x coefficient the standard error is 0.58. Now we are going to get the ANOVA table for this regression.

So this ANOVA table type this `print anova underscore lm fit1` we are getting this anova table for regression analysis what we are understanding here for independent variable is student population, so the degrees of freedom is one sum of square is 14200 when you divide this 14200 by 1 we are getting the mean sum of square. Then for a error term the degrees of freedom is 8. How it is 8 because there was a 10 data set so the degrees of freedom is $10 - 1$ - number of independent variables so $10 - 1$, 9 - 1 one independent variable so 8.

So the sum of square is 1530, so when you divide this 1532 by 8 we are getting the mean error sum of squares. So, F value is this 14200 by 191 you are getting this one so p value is very low this model is validated. Next what are you going to do we are going to draw the residual plot in x axis I have taken the student population that is independent variable in y axis this is not the sales it is the residual for the sales .

See, that next one we will see the studentized residual plot you run this, so this is your standardized residuals. So, we will plot this standardized residuals so this is a standardized visible so what is the interpretation from this is all the points are between + 2 and - 2 so we can say this assumption is valid. Next we will go for checking the normality of the error term. So now what is happening we are getting the qq plot when you run this code.

So this qq plot says that all the points are around this red line we can say this model is that is the assumption of the normality is tested it is correct. In every lectures you can follow this code you can verify this output I will also planning to share this code with you when you register this course. Now I will conclude what we have seen in this lecture in this lecture we have tested various assumptions about the regression models these assumptions we have tested with the help of different residual plots.

We have seen 4 types of residual plot 1 plot is a residuals against independent variable, the next one is the residual against our predicted values the third one is standardized the residual plot the fourth one is the normal probability plot. So, these different graphs helped us to test the assumption about the regression models. The next class we will discuss about the multiple regression models with some other examples, thank you.