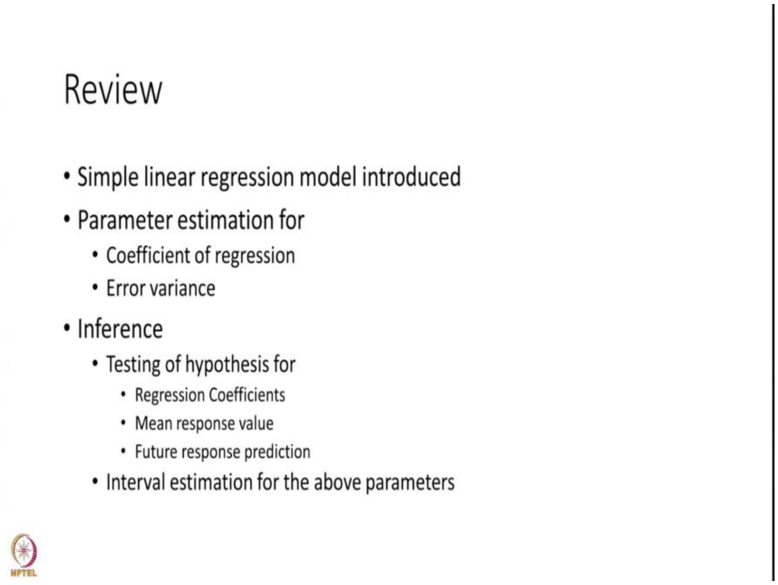
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor Hina A. Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 83
Regression Analysis - 3

Hello and welcome to the course on Dealing with Materials Data. We are going to continue our session from the previous two sessions on Regression Analysis. Let us quickly review what we have done in the past.

(Refer Slide Time: 00:32)



Review

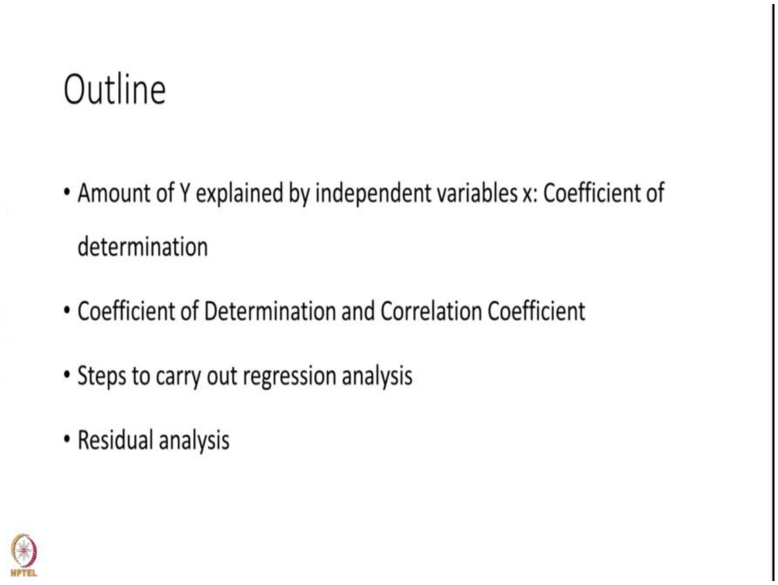
- Simple linear regression model introduced
- Parameter estimation for
  - Coefficient of regression
  - Error variance
- Inference
  - Testing of hypothesis for
    - Regression Coefficients
    - Mean response value
    - Future response prediction
  - Interval estimation for the above parameters

We introduced a simple linear regression model. We did the parameter estimations for coefficient of regression and also for error variance. Later on, we continued and we did some testing of hypothesis for regression coefficient, mean response value and future response prediction.

This process of hypothesis testing led us to have also the interval estimation of, all of, all the three of the above parameters discussed here. In the present case, we are going to now study in detail the analysis part of regression. There are three aspects to it, number one having done all this estimation and inference on the regression equation.

The question comes how much of Y your dependent variable or your response variable is explained by the dependent, independent variable X. This is decided through a coefficient of determination. Then we would like to know, because we already know that there exists something called a correlation between two variables X and Y. So, we would like to find out what is the relationship between coefficient of determination and correlation coefficient.

(Refer Slide Time: 02:14)

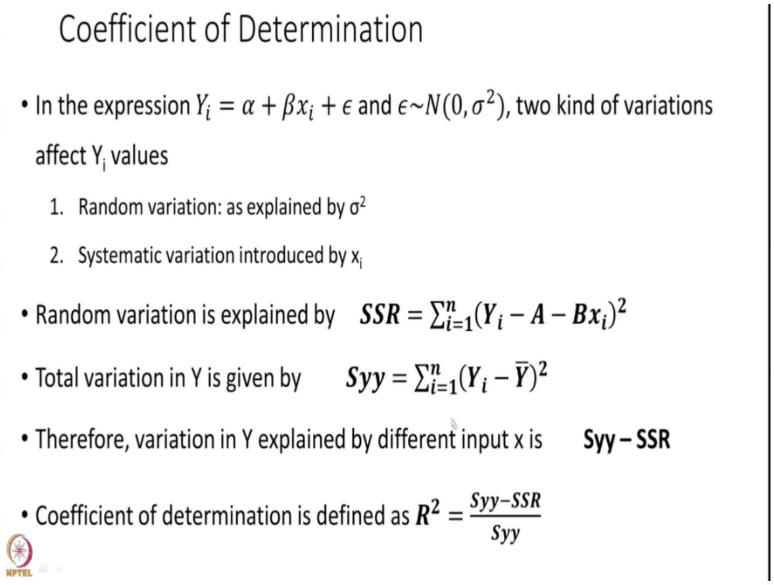


Outline

- Amount of Y explained by independent variables x: Coefficient of determination
- Coefficient of Determination and Correlation Coefficient
- Steps to carry out regression analysis
- Residual analysis

Next suppose you get a data. How do you go about carrying out the regression analysis? So we will give you steps how to go about doing it and having done it. The most important part is to know whether you are on the right path or not. That is you have done the regression analysis is the correct approach or not and this can be done through what is known as residual analysis.

(Refer Slide Time: 02:48)



Coefficient of Determination

- In the expression $Y_i = \alpha + \beta x_i + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$, two kind of variations affect Y_i values
  1. Random variation: as explained by σ^2
  2. Systematic variation introduced by x_i
- Random variation is explained by $\boldsymbol{SSR = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2}$
- Total variation in Y is given by $\boldsymbol{Syy = \sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- Therefore, variation in Y explained by different input x is **Syy – SSR**
- Coefficient of determination is defined as $\boldsymbol{R^2 = \frac{Syy-SSR}{Syy}}$

So let us begin. We start with coefficient of determination. You see when we write this expression

$$Y_i = \alpha + \beta x_i + \epsilon \text{ and } \epsilon \sim N(0,\ \sigma^2),$$

In this equation, we are trying to express actually two kinds of variations that effect the value of Y. One is the random variation, which is given by epsilon and it is mainly explained by the variance of the random error, which is sigma square and then we are talking about this independent variable or independent values xi that also introduced a systematic variation in Yi. So, the random variation we can say is estimated or explain by sum of squares of residual we have seen that in the past.

So, this sum of square of residuals can be expressed as

$$SSR = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2$$

The total variation in Y is given by, we denoted by Syy and it is given by

$$Syy = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

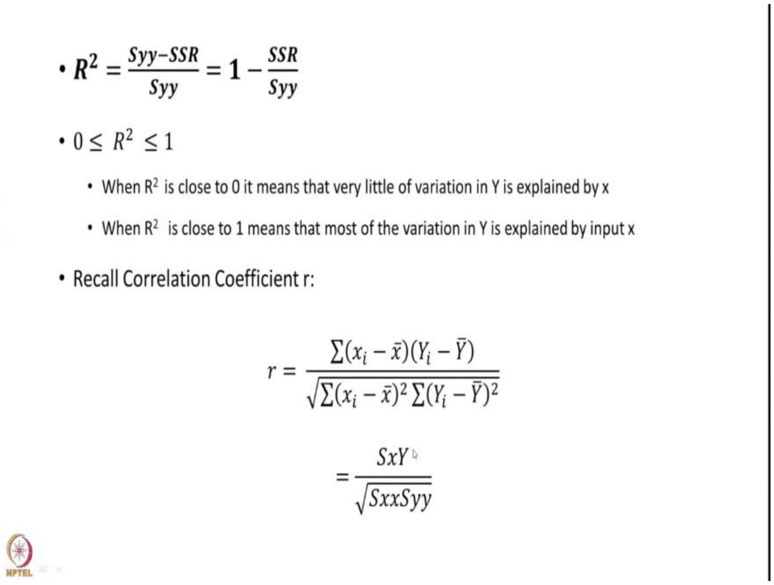
This is called the total variation. Now if this is the total variation and this is the variation which is introduced by sigma square that is the random variable.

Then the difference between the two inputs difference between the two sums of squares would show us or would should give us the variation that is explained by input variable x and therefore that is given by $Syy - SSR$.That is way total variation minus the variation due to residuals in that case coefficient of determination is defined as R square which is equal

$$\boldsymbol{R^2 = \frac{Syy - SSR}{Syy}}$$

In other words, we are trying to estimate the amount of variation cost in the total variation of Y value by the variable systematic variation introduced by xi.

(Refer Slide Time: 05:30)



- $\boldsymbol{R^2 = \frac{Syy-SSR}{Syy} = 1 - \frac{SSR}{Syy}}$
- $0 \le R^2 \le 1$
  - When R^2 is close to 0 it means that very little of variation in Y is explained by x
  - When R^2 is close to 1 means that most of the variation in Y is explained by input x
- Recall Correlation Coefficient r:

$$r = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(Y_i - \bar{Y})^2}}$$

$$= \frac{SxY}{\sqrt{SxxSyy}}$$

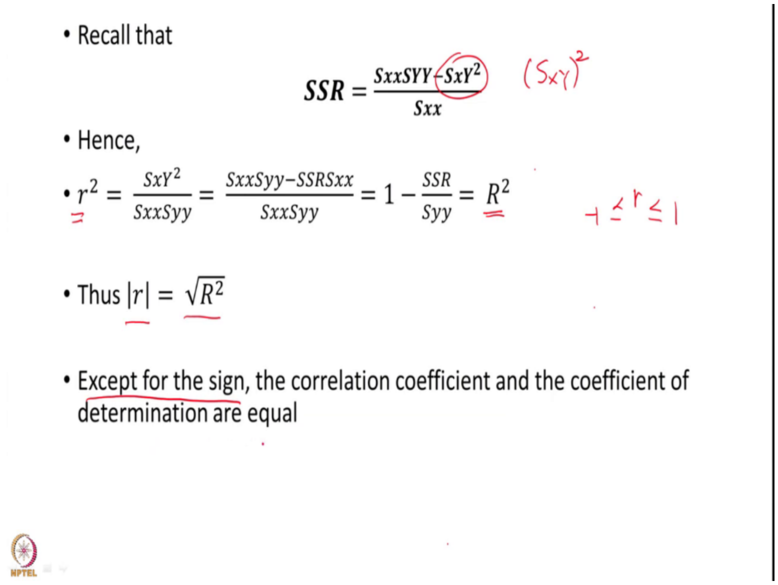
So, further if you write R square is equal to you can write it as 1 minus sums of squares of residuals divided by total variation in R, total variation in Yi and you can see that, this term is always positive and it has to be less than 1 and therefore the R square value lies between 0 and 1. Now when R square is close to 0, what does it imply? It implies that very little has been explained by the independent variable X, Y. When R square is close to 1, it means that most of the variation in Y is explained by input variable X.

Now having understood this, let us try to establish a relationship between correlation coefficient and the capital R square correlation coefficient we write it as r and capital R square. So, correlation coefficient is written as

$$r = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(Y_i - \bar{Y})^2}}$$

$$= \frac{SxY}{\sqrt{SxxSyy}}$$

(Refer Slide Time: 07:02)



- Recall that

$$SSR = \frac{SxxSYY - SxY^2}{Sxx} \qquad (S_{xY})^2$$

- Hence,
- $r^2 = \frac{SxY^2}{SxxSyy} = \frac{SxxSyy - SSRSxx}{SxxSyy} = 1 - \frac{SSR}{Syy} = R^2$ $\qquad -1 \le r \le 1$
- Thus $|r| = \sqrt{R^2}$
- Except for the sign, the correlation coefficient and the coefficient of determination are equal

So, this sums of squares of residual in those terms can be expressed as

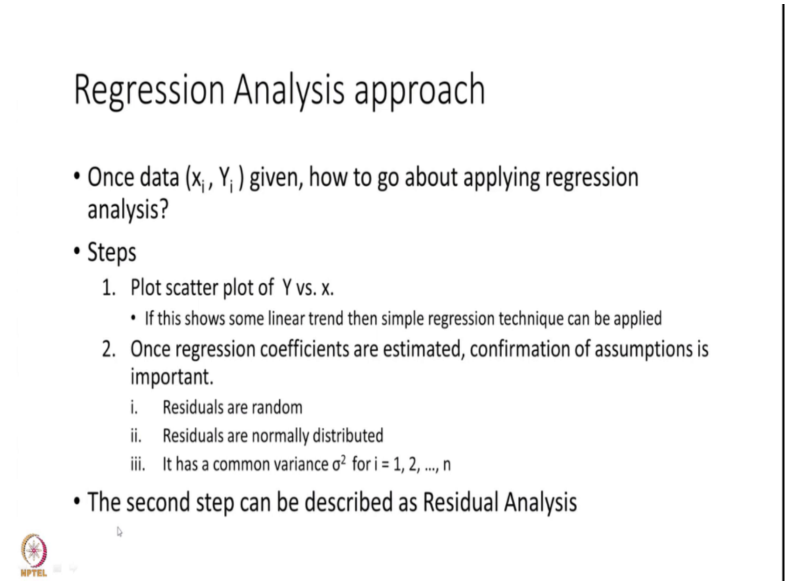
$$\boldsymbol{SSR = \frac{SxxSYY - SxY^2}{Sxx}}$$

- Hence,

$$r^2 = \frac{SxY^2}{SxxSyy} = \frac{SxxSyy - SSRSxx}{SxxSyy} = 1 - \frac{SSR}{Syy} = R^2$$

It will be shown that small r square is same as capital R square. But please remember that the small r lies between minus 1 and 1 and therefore we can only say that absolute value of r is a square root of R square. So, we can say that except for the sign, the correlation coefficient and the coefficient of determination are equal.

(Refer Slide Time: 08:25)



Regression Analysis approach

- Once data (x_i , Y_i) given, how to go about applying regression analysis?
- Steps
  1. Plot scatter plot of Y vs. x.
     - If this shows some linear trend then simple regression technique can be applied
  2. Once regression coefficients are estimated, confirmation of assumptions is important.
     - i. Residuals are random
     - ii. Residuals are normally distributed
     - iii. It has a common variance σ^2 for i = 1, 2, ..., n
- The second step can be described as Residual Analysis

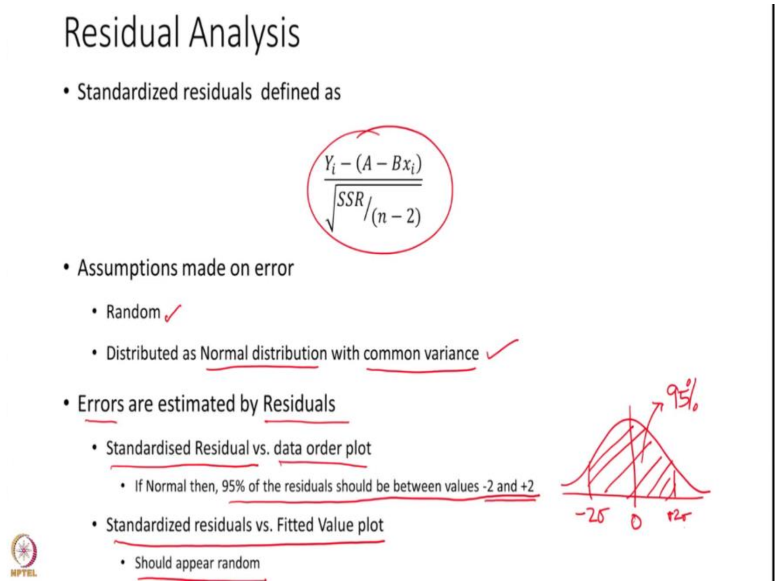
Now as I said having known the most of the parameters related to regression analysis and one comes across the data say Xi, Yi and then you have to apply regression analysis. How do you go about doing it? So, here we are explaining the steps to it. First it is important to have, we have to make I think we have moved forward.

So, we have the steps to follow number one. So, first step is plot a scatter plot of Y Verses X. If this shows some linear trend, you remember in the previous slides, we showed how you can estimate the linear trend exact trend verses approximate trend. So, if it shows some kind of linear trend. Then simple regression technique can be applied.

Once the regression coefficients are estimated confirmation of assumption is very important that is you have estimated the regression coefficient. You have conducted the hypothesis testing to make sure that none of the regression coefficient is 0. Then it is important that the whole process of regression analysis has been carried out under three very important assumptions.

The first assumption is that residuals are random. Second thing is residuals are normally distributed and the third is that they have a common value variance sigma square for all i equal to 1 2,3, etc up to n. So, now we will describe, because making a scatter plot we have already seen that in the previous slide, so there is nothing to show.

(Refer Slide Time: 11:06)



Residual Analysis

- Standardized residuals defined as

$$\frac{Y_i - (A - Bx_i)}{\sqrt{SSR/(n-2)}}$$

- Assumptions made on error
  - Random ✓
  - Distributed as Normal distribution with common variance ✓
- Errors are estimated by Residuals
  - Standardised Residual vs. data order plot
    - If Normal then, 95% of the residuals should be between values -2 and +2
  - Standardized residuals vs. Fitted Value plot
    - Should appear random

But we would like to go into the up second approach, which is called a residual analysis. First we defined a standardized residual for that. So, you remember the expression here is

$$\frac{Y_i - (A - Bx_i)}{\sqrt{SSR/(n-2)}}$$

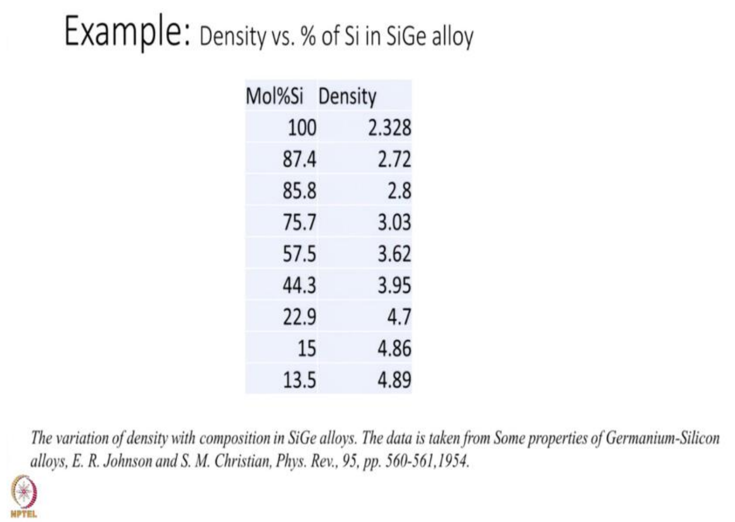
a residual and residuals divided by its standard deviation would be called a standardized residual and the standard deviation is I think by now it is very clear, it is sums of squares of residual divided by its degrees of freedom and we have to divide by taking the square root. Because this is the estimator for sigma square.

Now assumptions, we repeat assumptions made on the error or residual is that, the residuals are random. Let us see if the pen works again. So, first thing is that it is random. Second thing is that it is distributed normally and that it has a common variance. So, we know that the errors are estimated by residuals.

So, we have to first make a standardized residual. This is standardized residual verses data plot. If normal, then 95 percent of the residual should lie between minus 2 and plus 2. I hope you recall that if the distribution is normal and it is a standard normal distribution with mean 0. Then between minus, plus 2 sigma and minus 2 sigma. The data line or the probability of this area is 95 percent.

So, it means that when you take a standardized residual. They should lie, the sigma is we understand as 1. So, then it becomes it should lie between minus 2 and plus 2. Second thing we would like to have is that the standardized residuals verses fitted value plots and if this appears as a random scatter plot, then the randomness is also confirmed.

(Refer Slide Time: 13:48)



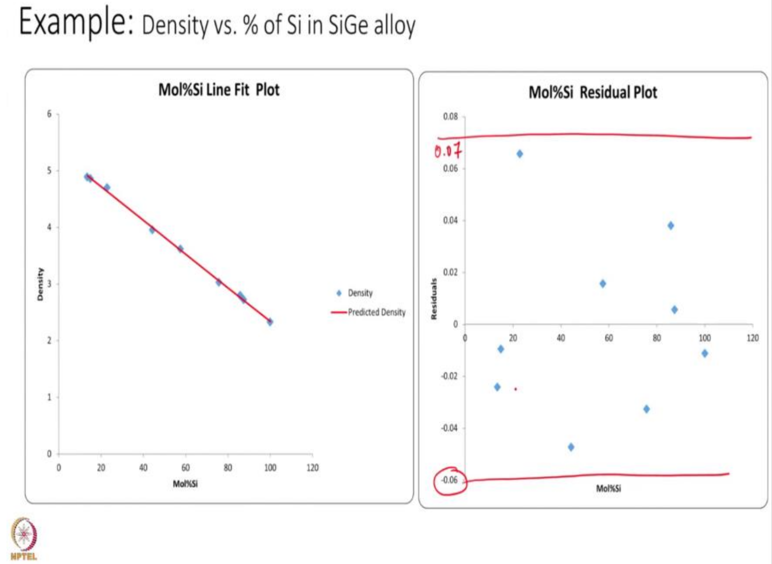
Example: Density vs. % of Si in SiGe alloy

Mol%Si	Density
100	2.328
87.4	2.72
85.8	2.8
75.7	3.03
57.5	3.62
44.3	3.95
22.9	4.7
15	4.86
13.5	4.89

The variation of density with composition in SiGe alloys. The data is taken from Some properties of Germanium-Silicon alloys, E. R. Johnson and S. M. Christian, Phys. Rev., 95, pp. 560-561,1954.

Let us take an example, here I have an example of germanium silicon alloys in which the density of the alloy is tested against the percentage of silicon in the alloy.

(Refer Slide Time: 14:03)



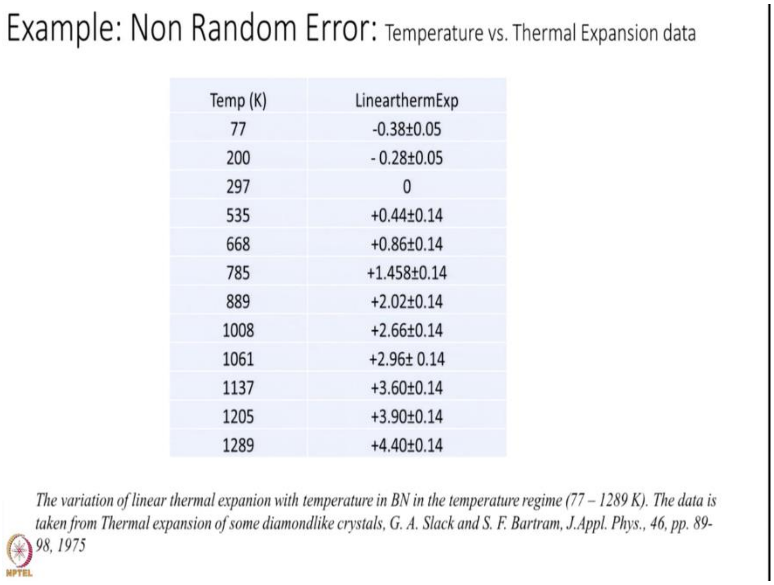
Example: Density vs. % of Si in SiGe alloy

Now I have not shown you the scatter plot, but you see that this data points. The blue colour are actually density data points. This is the, on the x axis we have amount of silicon in the alloy and y axis has the density and then these points are the actual points of density as measured when silicon value is given on the x axis.

It is pretty must in straight line and therefore it is perfectly fine to apply the linear model. Now that we have applied the linear model. There is a line fit and now you look at the residual plots and in this residual plots again, it is the, we have plotted against the percentage of silicone against the standardized residual and please note that the data is completely data is completely between this is minus 0.06 to about plus 0.07.

It means that it is normal because most of the data is between the minus 2 and plus 2 limit. It is actually in a very short, in a very narrow spam and you can see that the plot is quite scattered as you can see it is completely a scattered plot randomly scattered. So, our assumption of randomness is also justified.

(Refer Slide Time: 15:54)



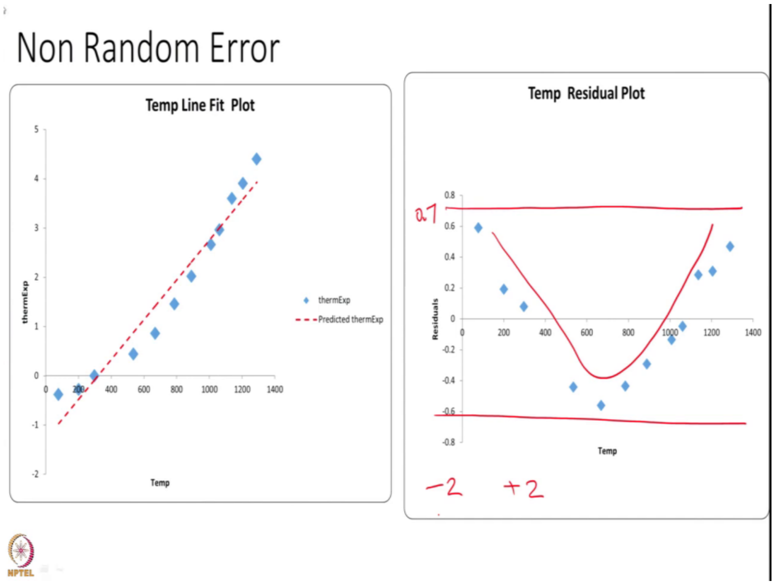
Example: Non Random Error: Temperature vs. Thermal Expansion data

Temp (K)	LinearthermExp
77	-0.38±0.05
200	- 0.28±0.05
297	0
535	+0.44±0.14
668	+0.86±0.14
785	+1.458±0.14
889	+2.02±0.14
1008	+2.66±0.14
1061	+2.96± 0.14
1137	+3.60±0.14
1205	+3.90±0.14
1289	+4.40±0.14

The variation of linear thermal expanion with temperature in BN in the temperature regime (77 – 1289 K). The data is taken from Thermal expansion of some diamondlike crystals, G. A. Slack and S. F. Bartram, J.Appl. Phys., 46, pp. 89-98, 1975

Let us take another example. This is an example where I want to show, what do you see if the error not random? So, here we have a data on linear thermal expansion in the alloy and the thermal expansion along with the temperature is shown here. The plot we have made is without considering this data error values given. So, we have plotted like 77 against minus 0.388 etc, etc. The temperature is taken in kelvin.

(Refer Slide Time: 16:37)



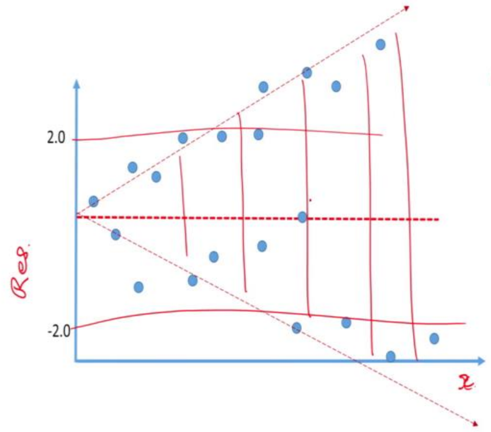
Non Random Error

Let us move on. You see this, the actual scatter plot of the data does not appear to be linear. But well it is not very far from the linear either. So, if we actually plot the line it comes like this. This is a least squares line. So, this is the regression line we are plotting. But if you look at these residual plot against it is temperature verses standardized residual. Though the residuals are well within the limit, please note.

Let us put up the pen again, so please note that. This is 0.7 and here it is minus 0.6. So, the data is well within minus 2 and plus 2 regime. But it is not scattered in a random fashion. There is a pattern to it. There is a pattern to it and this pattern is reflected in here. This pattern is reflected in here. Therefore, this very clearly says that number one, the randomness is not justified and therefore linear regression simple linear regression is not a correct model choice to express your thermal expansion in terms of temperature.

(Refer Slide Time: 18:23)

Non-constant variance (Hetroscadasticity)



σ^2 common

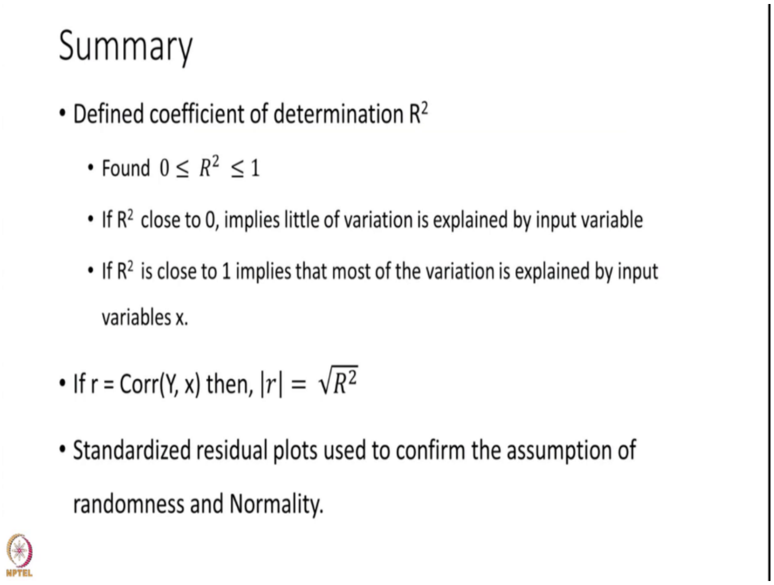
$i=1,2,\ldots,n$

incorrect

I could not get any other example where the random error gets distributed which shows that it is the variance is not constant. So, I have artificially generated a graph, in this basically what we are trying to show, is that, in such situations, not even that the error have gone beyond 2 plus 2 and minus 2 that is one aspect.

But before that we can also see that the error seems to be growing with the x axis. This is x axis and this is residuals. Then you can see that the value, the variation in residual tends to increase. This is called heteroscedasticity. It means that sigma square which you have assumed to be same common, for i is equal to 1 to n is incorrect. It does not apply here; sigma square actually is changing. So, this is called heteroscedasticity a plot can look something like this.

(Refer Slide Time: 19:43)



Summary

- Defined coefficient of determination R^2
  - Found $0 \leq R^2 \leq 1$
  - If R^2 close to 0, implies little of variation is explained by input variable
  - If R^2 is close to 1 implies that most of the variation is explained by input variables x.
- If r = Corr(Y, x) then, $|r| = \sqrt{R^2}$
- Standardized residual plots used to confirm the assumption of randomness and Normality.

So, now to summarize what we have learned today. We first defined the coefficient of determination R square. It is actually shows the variation, due to rather variation explained by the independent variable X in Y, compared to the total variation that has been introduced in Y that is the total variation of Y what percentage of variation is introduced by the variation in X.

If R square is close to 0, it implies that very little variation is explained by input variable. If R square is close to 1, it implies that most of the variation is explained by the input variable X. The correlation coefficient and the coefficient of determination are related. Correlation coefficient square is coefficient of determination or the absolute value of correlation coefficient is equal to square root of coefficient of determination. We have seen that standardized residual plots can be used to confirm the assumption of randomness normality and I have forgotten to mention. Let us mention it here and common variance of epsilon, thank you.