

Dealing with Material Data: Collection, Analysis and Interpretation
Professor M P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 75
Graphical Handling of Data

(Refer Slide Time: 00:26)

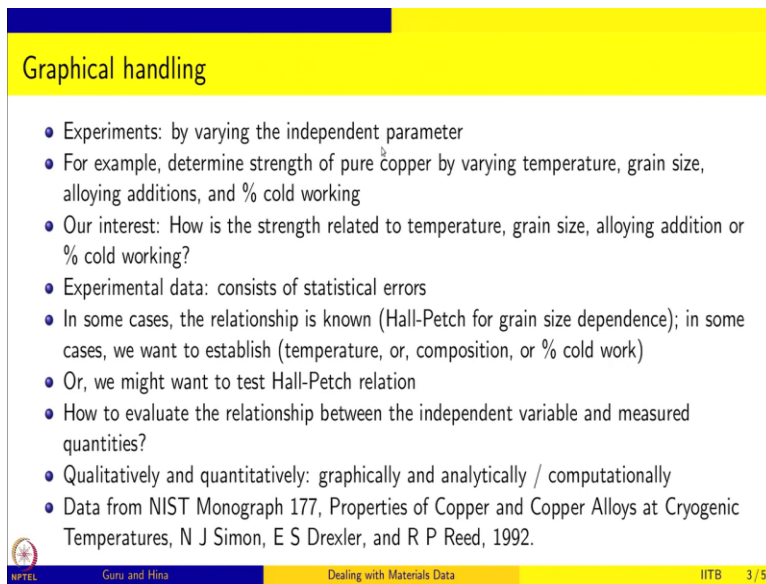
Module 5: Fitting and graphical handling

Graphical handling of data

The slide features a yellow header bar with the text "Module 5: Fitting and graphical handling". Below this, the main content area is white with the text "Graphical handling of data". At the bottom, there is a footer bar with a blue section on the left containing the NPTEL logo and the text "Guru and Hina", and a yellow section on the right containing the text "Dealing with Materials Data" and "IITB 2 / 5".

Welcome to Dealing with Materials Data, this is a course on Collection Analysis and Interpretation of Data from Material Science and Engineering. We are in module 5 which is on fitting and graphical handling of data. In this session I am going to be talking primarily about graphical handling, but we will also do a little bit of fitting along the way.

(Refer Slide Time: 00:40)



Graphical handling

- Experiments: by varying the independent parameter
- For example, determine strength of pure copper by varying temperature, grain size, alloying additions, and % cold working
- Our interest: How is the strength related to temperature, grain size, alloying addition or % cold working?
- Experimental data: consists of statistical errors
- In some cases, the relationship is known (Hall-Petch for grain size dependence); in some cases, we want to establish (temperature, or, composition, or % cold work)
- Or, we might want to test Hall-Petch relation
- How to evaluate the relationship between the independent variable and measured quantities?
- Qualitatively and quantitatively: graphically and analytically / computationally
- Data from NIST Monograph 177, Properties of Copper and Copper Alloys at Cryogenic Temperatures, N J Simon, E S Drexler, and R P Reed, 1992.

IPTEL Guru and Hina Dealing with Materials Data IITB 3 / 5

We have seen that if you have a set of data, you can calculate the average and the distribution of the data and from which you can estimate the mean and standard deviation and if you know something about the probability distribution from which this data is sampled you can even get properties of that probability distribution by looking at the data. But those are very simple things that we have done, so either parameter estimation as a point estimate or as an interval to say that this is the probability that the true mean will lie in this range and things like that.

Now, what we are interested is in a slightly more detailed analysis, because when we do experiments typically we vary some independent parameter and make some measurement. For example, you can say that, I will change the grain size and I will look at the strength, or you can say I will change the composition, I will look at the lattice parameter, or you can say I will change the temperature and I will look at the thermal expansion coefficient.

So, these are the kind of things that we do and the composition or temperature or grain size, then becomes the independent parameter and the measurement that you make in terms of strength or lattice parameter or thermal expansion coefficient, then becomes the dependent variable.

So, here is an example. So, determine strength of pure copper, you can do that by varying temperature, grain size, alloying addition or percentage cold working. Now, our interest is typically in knowing how strength is related to these parameters, but we cannot directly read it off from the data because the data consists of statistical errors.

In some cases for example between strength and grain size, you might think that I know what is the relationship, Hall-Petch is the relationship. In some cases we might not know, how does it change with temperature or composition or percentage cold work maybe they have no idea, or somebody says that the relationship is Hall-Petch, I am doing the experiment to actually test whether it is Hall-Petch.

So, in all these cases we have to establish a relationship between independent parameter and the measurement that we are making and there are several scenarios where the relationship is known or where you are trying to test that relationship or where you are just exploring and trying to find if there is a relationship and so on and so forth.

So, we want to do this quantitatively as well as qualitatively, so you can do it graphically and analytically or computationally. And from NIST Monograph 177, properties of copper and copper alloys at cryogenic temperatures, there is a whole bunch of data on all these aspects which we will be using in this session extensively.

(Refer Slide Time: 03:42)

Functional relationship

- Independent variable x_i
- Measurement y_i
- $y_i = f(x_i; \theta_k)$ where θ_k are the m unknown parameters


NPTEL Guru and Hina Dealing with Materials Data IITB 4/5

And to describe what is it that we are trying to do, so there are independent variables x_i and we are making measurement y_i and y_i is related to x_i , but there might also be other unknown parameters that go into this expression. And so we need to estimate these and we need to establish this relationship, so that is what we are trying to do.

(Refer Slide Time: 04:02)

Plot

- Plot the data: always good to identify the trends
- Plot with error bars
- Once you see the trend, we can fit
- Good idea to plot the data and fit together
- Good idea to analyse the residuals
- Ideally, residuals should be random!
- First example: The variation of density and lattice parameter 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.
- Second example: The variation of linear thermal expansion 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.



NPTEL Guru and Hina Dealing with Materials Data IITB 5/5

So, the first thing to do is always to plot the data and just looking at the data it is possible to identify trends and if the error bars are available, you should always plot with error bars, because without error bars sometimes you might read of wrong trends, once you see the trend then you can fit. And every time you do fitting it is always a good idea to plot the data and the fit together. And it is also a good idea to analyze the residuals, what is the residual?

So, if you have made a fit and if you have data, how far are these data points from your fit is a good thing to look at and if it is proper fit, because the residual is an error, it should be a random error, which means it should show the characteristics of being a random variable. Now, before we go into the copper database we are going to do some very simple analysis just to show how plotting is a good idea and after plotting how do we go about fitting the data in R.

So, I am going to use 2 data sets, one is variation of density and lattice parameter with composition in silicon germanium alloys, this is because we know that with composition the lattice parameter will change linearly and this is known as Vegard's law, so we are going to check that and get the Vegard's law coefficients and that is this exercise.

The second one is a variation of linear thermal expansion with temperature in boron nitrite in a given temperature range 77 to 1289 Kelvin, this is also a data taken from the literature and in this case we are going to do a little bit more exploratory than this, because in this case it is known that lattice parameter and composition should be straight line, so if you plot and you see that it is a

straight line, then you can go ahead fit and get the parameters, but here let us say that we do not know what the relationship is and we are going to explore and find out how it is done. So, these are the 2 exercises that we are going to do.

(Refer Slide Time: 06:30)

Module: Data processing using R

M P Gururajan and Hina A Gokhale

Indian Institute of Technology Bombay, Mumbai

1 Interval estimation and confidence levels

```
Y <- read.csv("../Data/SiGeLatticeParameterDensity.csv")
c <- Y$Mol.percent.Si
a <- Y$Lattice.constant
plot(c,a)
fit <- lm(a~c)
abline(fit$coefficients)
plot(fit$residuals)
qqnorm(fit$residuals)
```

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

```
> Y <- read.csv("Data/SiGeLatticeParameterDensity.csv")
> c <- Y$Mol.percent.Si
> a <- Y$Lattice.constant
> plot(c,a)
> fit <- lm(a~c)
> abline(fit$coefficients)
> plot(fit)
```

Variable	Value
fit	List of 12
Y	11 obs. of 3 variables
a	num [1:11] 5.43 5.45 5.46 5.47 5.52 ...
c	num [1:11] 100 87.4 85.8 75.7 57.5 4...

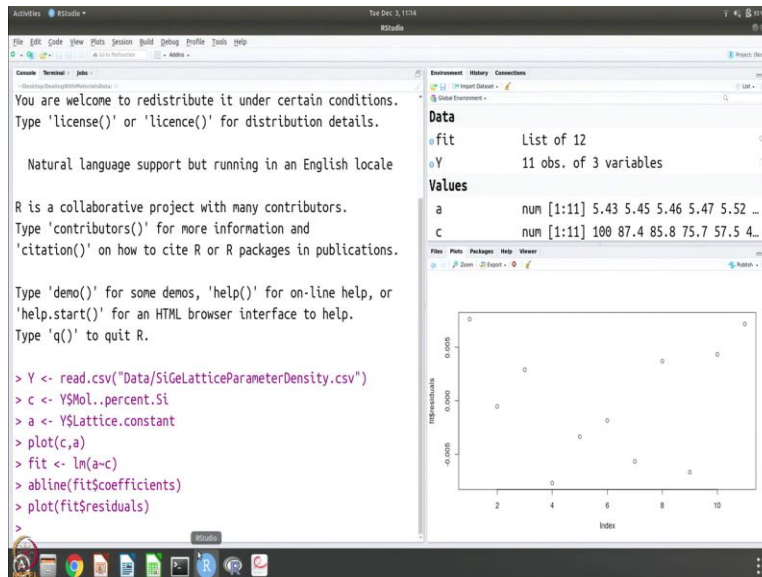
Plot showing a linear relationship between c (x-axis) and a (y-axis). The x-axis ranges from 0 to 100, and the y-axis ranges from 0.45 to 0.65. The data points are approximately: (100, 0.43), (87.4, 0.45), (85.8, 0.46), (75.7, 0.47), (57.5, 0.52), (40, 0.55), (30, 0.58), (20, 0.60), (10, 0.62), (5, 0.64).

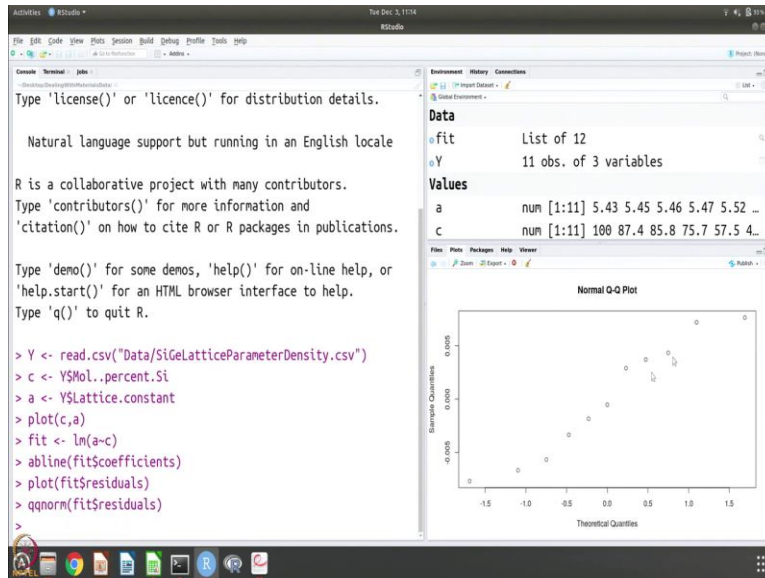
So, let us start R and let us do the first thing. So, we are going to first read the data, silicon, germanium, lattice parameter and density and in composition I am going to store the mole percent of silicon and in the variable a, I am going to store the lattice constant and let us say plot c and a, so let us do this as the first exercise.

So, we have plotted and you can see that as composition changes the lattice parameter changes and from the plot it is clear that it is a straight line. So, then we can ask a straight line fit to be made and that is done using the command, so we that is done using the command fit, we have mount of linear fit between a and c and we are going to plot a line using the fit coefficients.

So, these two I am going to do right now, so let us do this. So, we fitted and so it is as simple as that just say fit and you can look at the coefficients of the fitted line and then draw a line and you can see that we have a nice line running through the points and you can see that the error is random the points lying on either side but quite close to the data and so that is what is done, you can say that you can plot the fit, residuals.

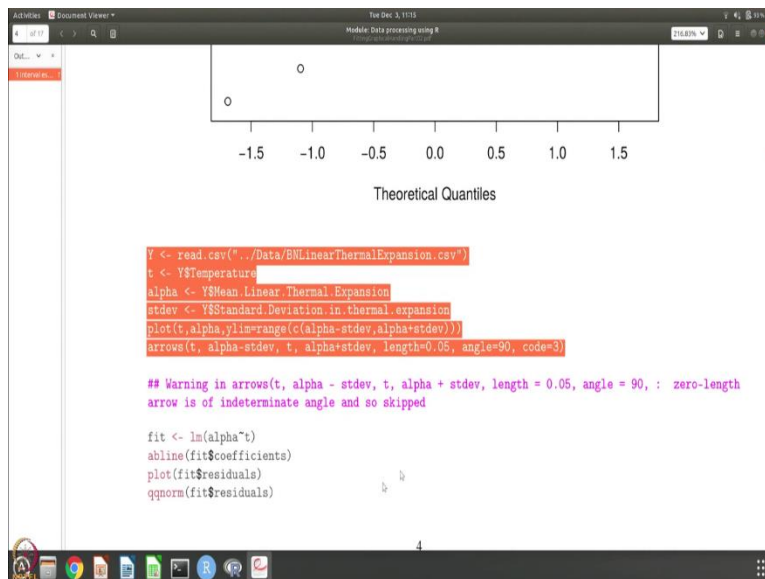
(Refer Slide Time: 08:24)

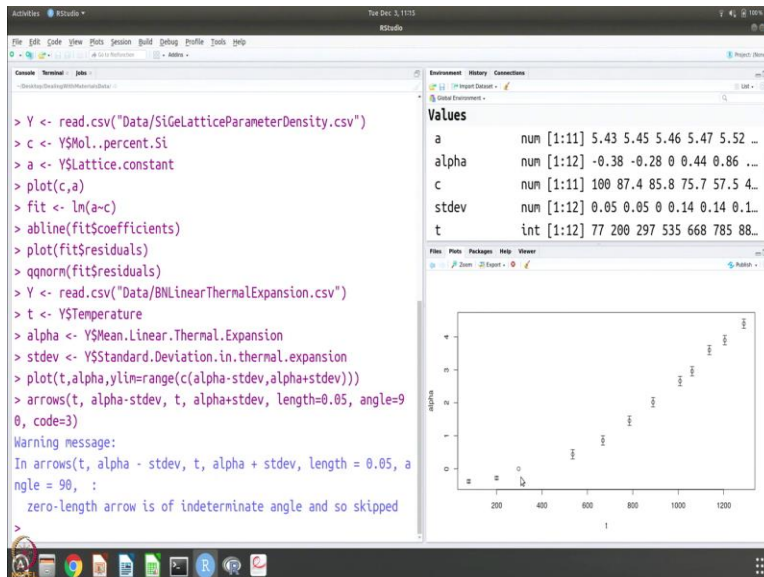
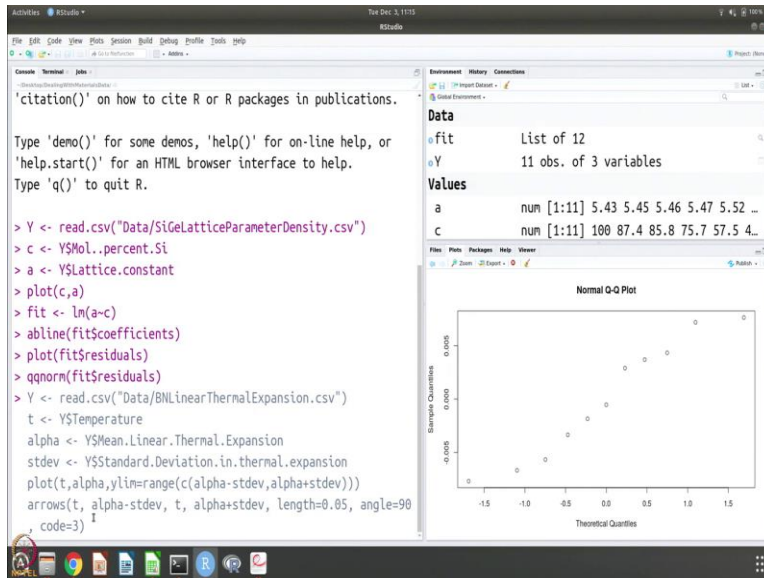




So, this is the error you can see about 0 on either side the data is scattered very nicely and you can also plot the qqnorm of the residuals to know whether that is normal or not, remember this is another way of knowing if your data is distributed normally, so if it is more or less a straight line, you know that the fit the error is normally distributed. So, this is a rather straightforward exercise.

(Refer Slide Time: 09:12)

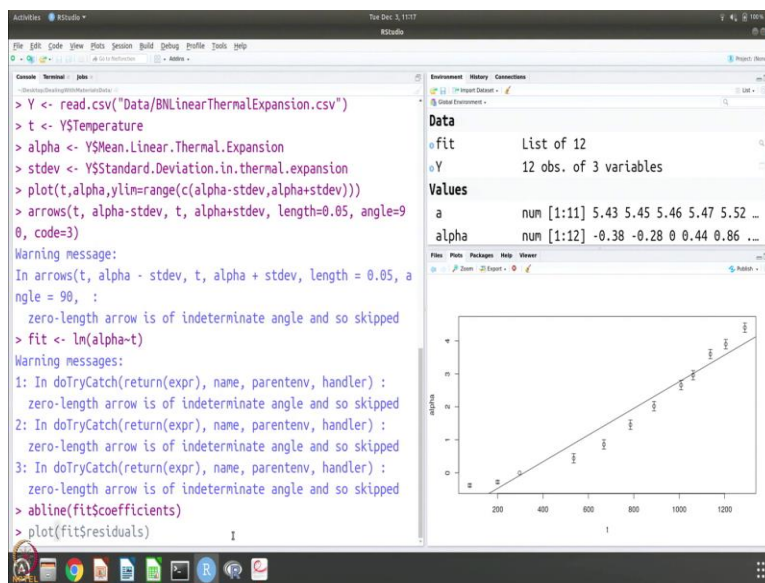
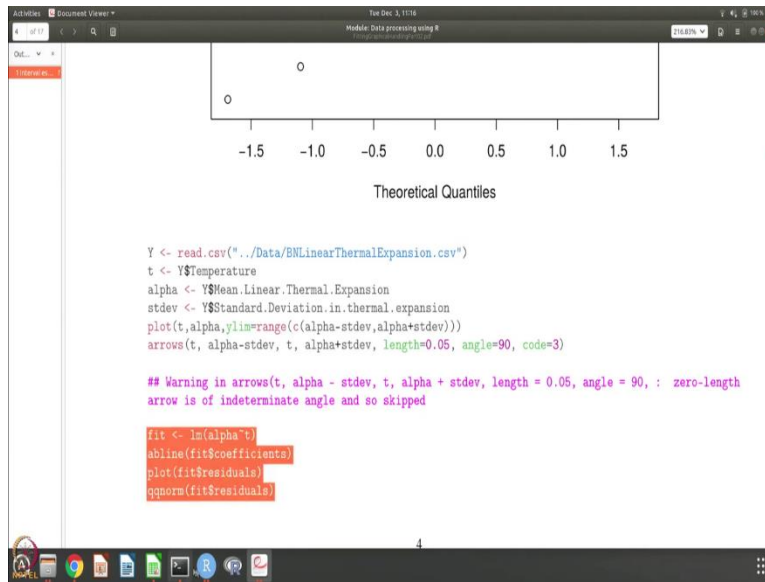


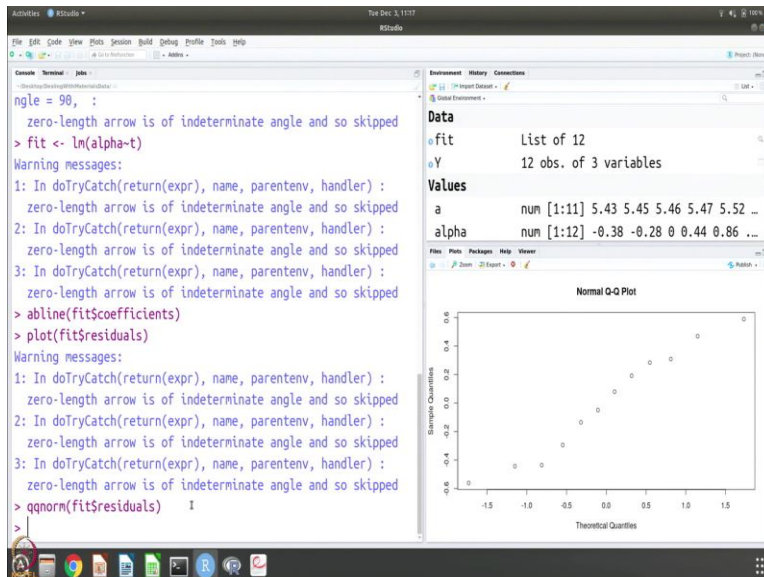
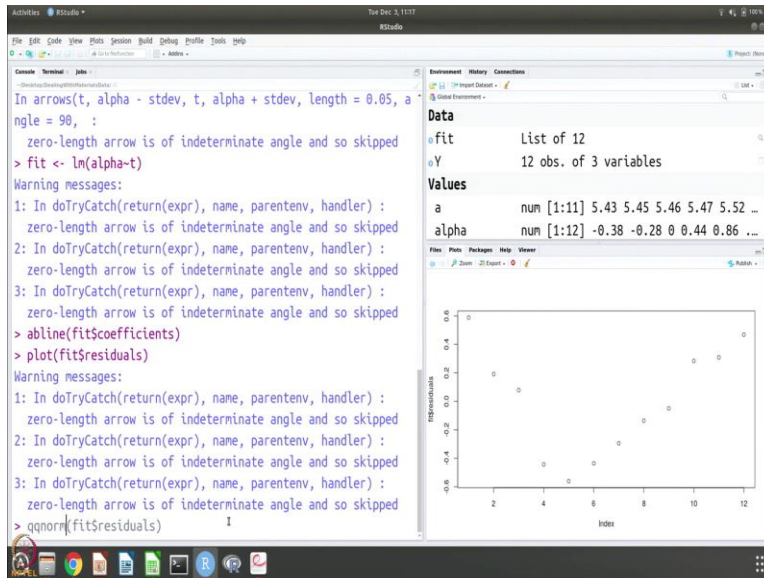


Let us, go now to the next one which is to get the thermal expansion coefficient of boron nitride. So, in this case again, we read the data BNlinear thermal expansion and restore temperature as t and we store the linear thermal expansion as alpha and in this case, there is also error that is given standard deviation, so we store that as standard deviation.

Then we plot the data and then we put the error bars and that is what is done using this, so let us do this and it is telling that 0 length arrow that is because one of the data points is the reference with respect to which everything is taken, so it has no error bar or error bar is 0 that is what it is telling. Now, by looking at this data, one can see that this could be a straight line and this could be a straight line. So, there could be two straight lines for this data.

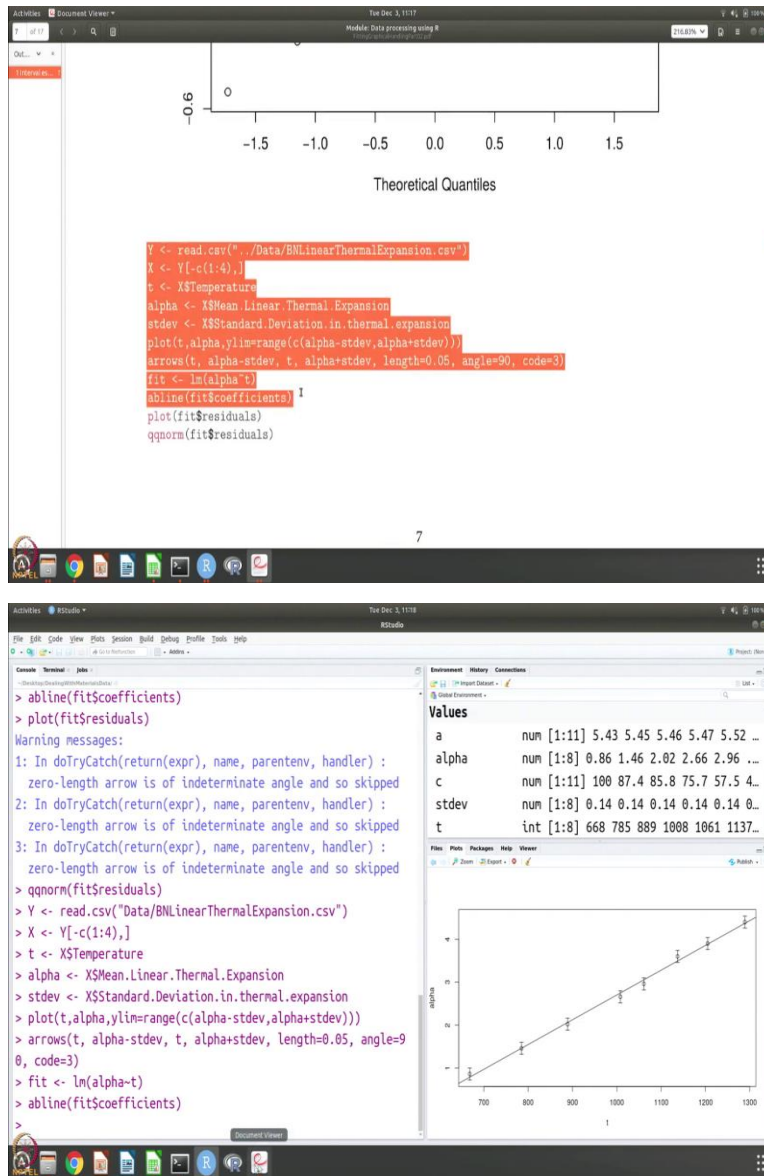
(Refer Slide Time: 10: 27)





So, if we do not do that, if you just fit, let us do that. So, you can see that you can fit a straight line, but the straight line is not looking like a very good fit for this data and it would be easier or better if we can fit a straight line here and if we can fit a straight line here. So, this is also clear if you look at the fit residuals, so residuals actually show a trend. And you can also do the qqnorm of the residuals and you can see that that also is not quite a straight line.

(Refer Slide Time: 11:47)



So, one thing to do is that let us take the data points, let us leave out the first four and take the remaining data points and do a fit, so I am going to that first, so what we have done is we take the same data, now we leave out the first four data points and then for the rest of them we try to fit a linear curve and you can see that it fits very nicely and here the data is above and below and probably normally distributed.

(Refer Slide Time: 12:33)

Theoretical Quantiles

```
Y <- read.csv("Data/BNLinearThermalExpansion.csv")

## Warning in file(file, "rt"): cannot open file 'Data/BNLinearThermalExpansion.csv': No
such file or directory
## Error in file(file, "rt"): cannot open the connection

X <- Y[,c(5:12),]
t <- X$Temperature
alpha <- X$Mean.Linear.Thermal.Expansion
stdev <- X$Standard.Deviation.in.thermal.expansion
plot(t,alpha,ylim=range(c(alpha-stdev,alpha+stdev)))
arrows(t, alpha-stdev, t, alpha+stdev, length=0.05, angle=90, code=3)

## Warning in arrows(t, alpha - stdev, t, alpha + stdev, length = 0.05, angle = 90, : zero-length
```

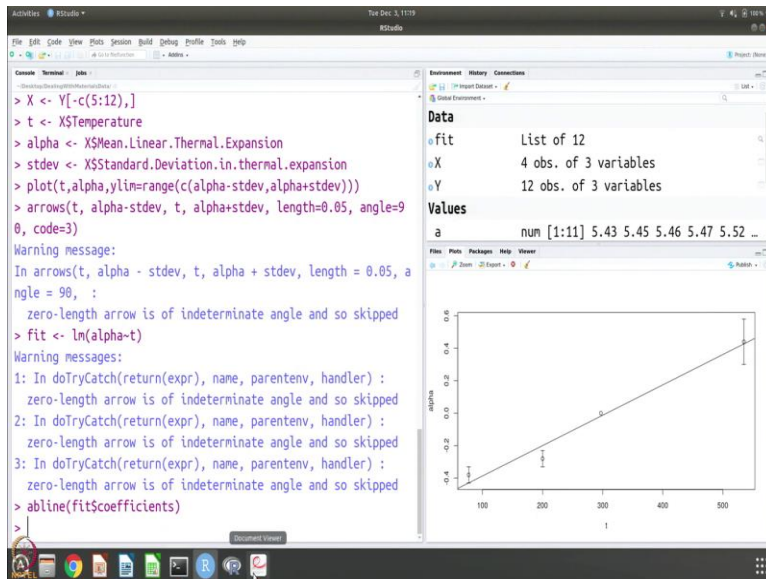
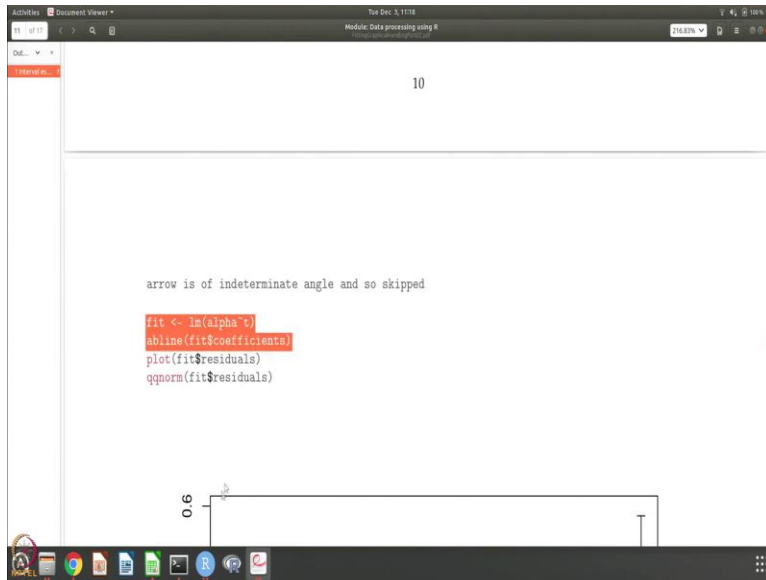
10

```
zero-length arrow is of indeterminate angle and so skipped
3: In doTryCatch(return(expr), name, parentenv, handler) :
zero-length arrow is of indeterminate angle and so skipped
> qqnorm(fit$residuals)
> Y <- read.csv("Data/BNLinearThermalExpansion.csv")
> X <- Y[,c(1:4),]
> t <- X$Temperature
> alpha <- X$Mean.Linear.Thermal.Expansion
> stdev <- X$Standard.Deviation.in.thermal.expansion
> plot(t,alpha,ylim=range(c(alpha-stdev,alpha+stdev)))
> arrows(t, alpha-stdev, t, alpha+stdev, length=0.05, angle=90,
code=3)
> fit <- lm(alpha~t)
> abline(fit$coefficients)
> X <- Y[,c(5:12),]
t <- X$Temperature
alpha <- X$Mean.Linear.Thermal.Expansion
stdev <- X$Standard.Deviation.in.thermal.expansion
plot(t,alpha,ylim=range(c(alpha-stdev,alpha+stdev)))
arrows(t, alpha-stdev, t, alpha+stdev, length=0.05, angle=90,
code=3)
```

Environment: Global Environment

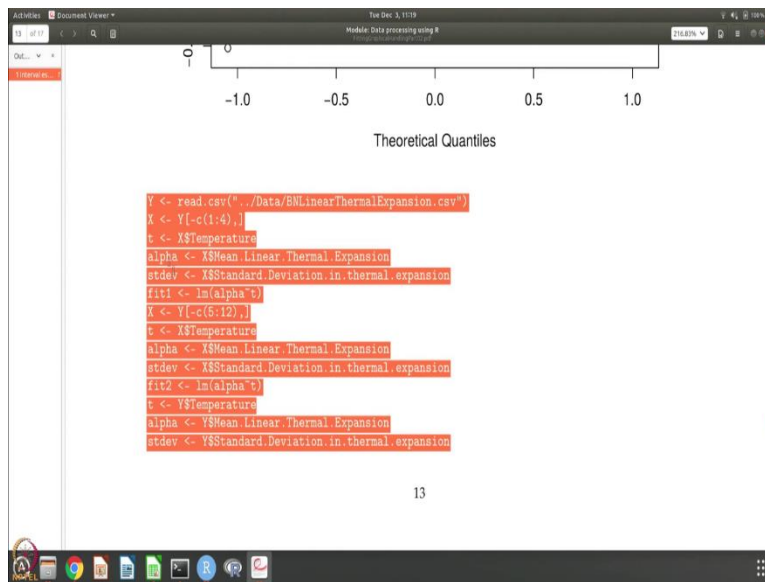
Values	
a	num [1:11] 5.43 5.45 5.46 5.47 5.52 ...
alpha	num [1:8] 0.86 1.46 2.02 2.66 2.96 ...
c	num [1:11] 100 87.4 85.8 75.7 57.5 4...
stdev	num [1:8] 0.14 0.14 0.14 0.14 0.14 0...
t	int [1:8] 668 785 889 1008 1061 1137...

The plot shows a linear relationship between temperature (t) on the x-axis and the mean linear thermal expansion (alpha) on the y-axis. The x-axis ranges from 700 to 1300, and the y-axis ranges from 0 to 4. Data points are plotted with vertical error bars, and a solid black line represents the linear fit. The data points are approximately at (668, 0.86), (785, 1.46), (889, 2.02), (1008, 2.66), (1061, 2.96), (1137, 3.16), (1200, 3.36), and (1260, 3.56).



You can do a similar thing for the first four data points, how do we do that? So, we can do this, so you can leave out from 5 to 12 and then do the plot and of course also do the fitting and so you can see that it also fits for these two data points.

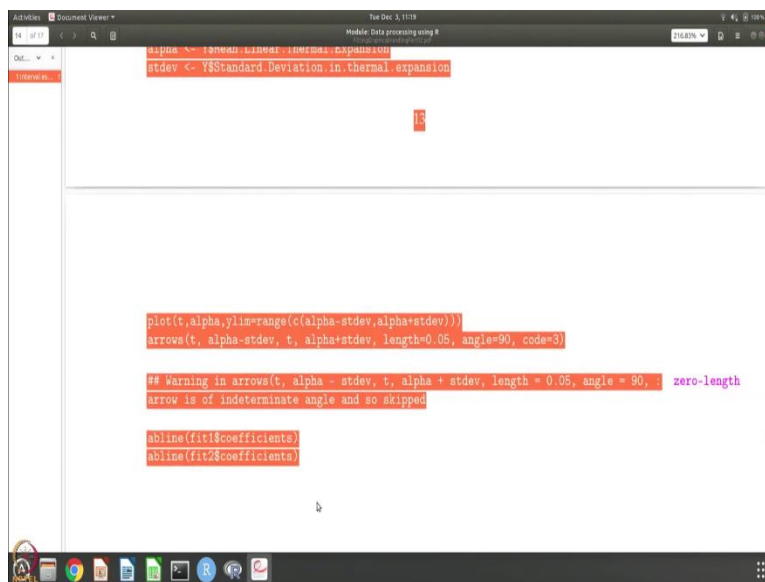
(Refer Slide Time: 13:24)



The screenshot shows the RStudio interface. At the top, there is a plot titled "Theoretical Quantiles" with a horizontal axis ranging from -1.0 to 1.0. Below the plot, there is a block of R code that reads data from a CSV file, calculates the mean and standard deviation of the 'alpha' variable, and fits a linear model. The code is as follows:

```
Y <- read.csv("../Data/BNLinearThermalExpansion.csv")
X <- Y[,c(1:4)]
t <- X$Temperature
alpha <- X$Mean.Linear.Thermal.Expansion
stdev <- X$Standard.Deviation.in.thermal.expansion
fit1 <- lm(alpha~t)
X <- Y[,c(5:12)]
t <- X$Temperature
alpha <- X$Mean.Linear.Thermal.Expansion
stdev <- X$Standard.Deviation.in.thermal.expansion
fit2 <- lm(alpha~t)
t <- Y$Temperature
alpha <- Y$Mean.Linear.Thermal.Expansion
stdev <- Y$Standard.Deviation.in.thermal.expansion
```

The number 13 is displayed at the bottom of the R console area.



The screenshot shows the RStudio interface with R code for plotting and adding annotations. The code is as follows:

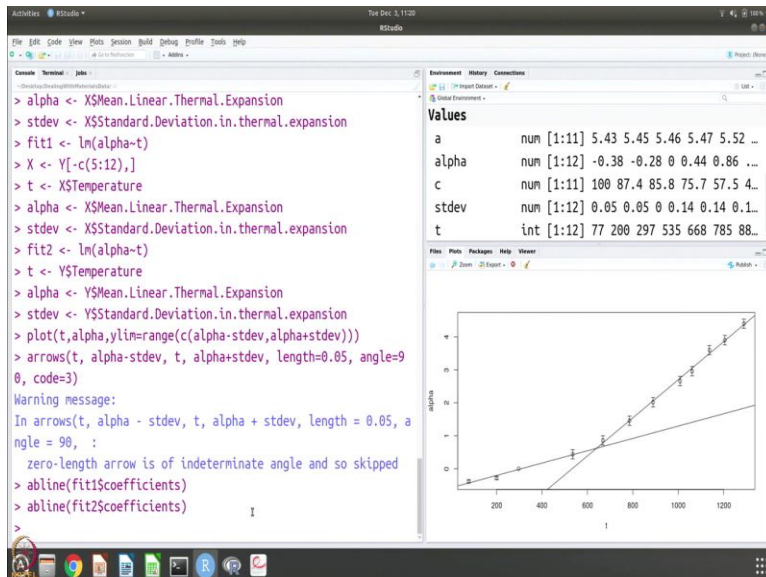
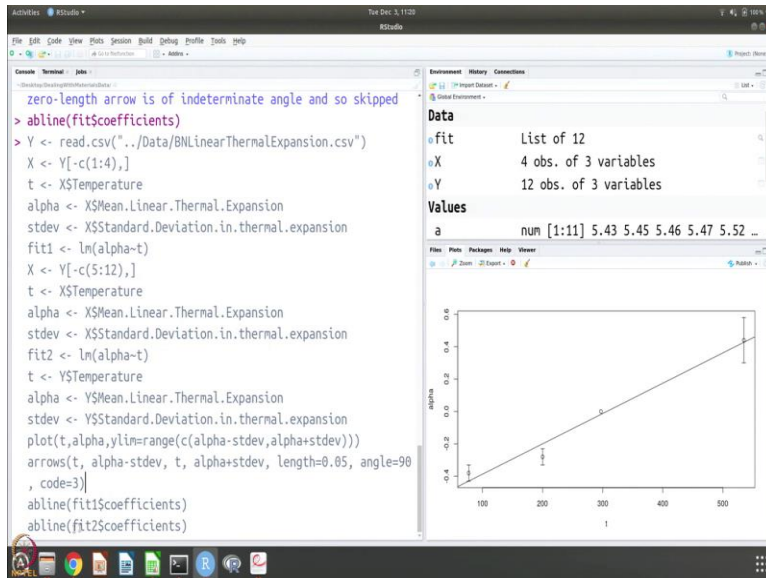
```
alpha <- X$Mean.Linear.Thermal.Expansion
stdev <- Y$Standard.Deviation.in.thermal.expansion

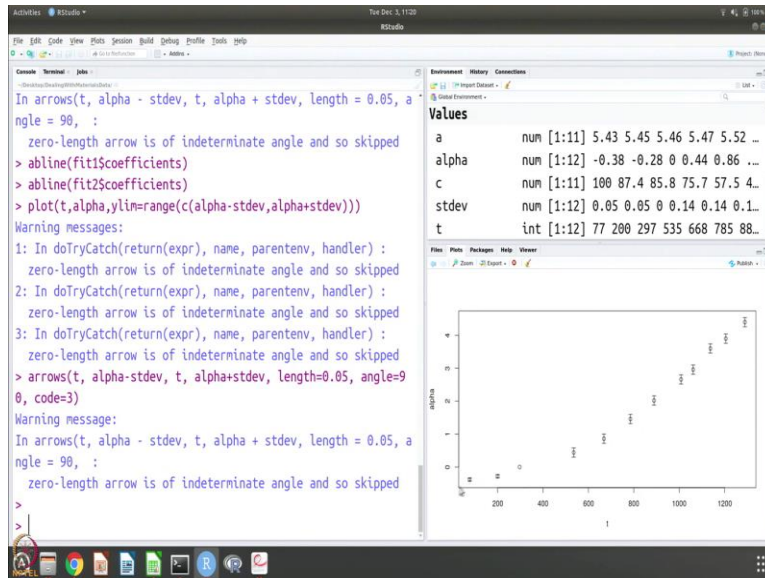
plot(t, alpha, ylim=range(c(alpha-stdev, alpha+stdev)))
arrows(t, alpha-stdev, t, alpha+stdev, length=0.05, angle=90, code=3)

## Warning in arrows(t, alpha - stdev, t, alpha + stdev, length = 0.05, angle = 90, : zero-length
arrow is of indeterminate angle and so skipped

abline(fit1$coefficients)
abline(fit2$coefficients)
```

The number 13 is displayed at the bottom of the R console area.





And of course we can do the other thing namely that we can take the data to both the fits, so I am going to do this, so what is it that we are doing? We are reading the data and we are leaving out the first to four data points we are making one fit, we are leaving out the 5 to 12 data points we making the second fit, we are going to plot the data and we are going to plot the both the fit lines. So, you can see that, so this is fit for a straight line and this fit is for a straight line, so you can say that there are two regimes below 600, for example, you can fit it with one linear curve and above 600 you can fit it with another linear curve. But if you just look at the data, let us do that, it is also possible that the data looks like a parabolic curve you might think that this might fit a nice parabola, is that true? Well one can check, and that is what the last exercise now is.

(Refer Slide Time: 15:08)

```
Y <- read.csv("../Data/BNLinearThermalExpansion.csv")
t <- Y$Temperature
alpha <- Y$Mean.Linear.Thermal.Expansion
stdev <- Y$Standard.Deviation.in.thermal.expansion

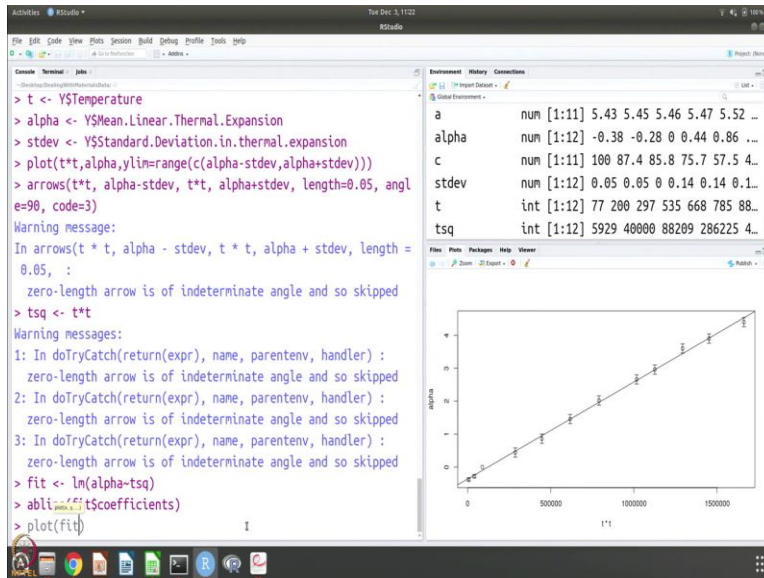
plot(t*t, alpha, ylim=range(c(alpha-stdev, alpha+stdev)))
arrows(t*t, alpha-stdev, t*t, alpha+stdev, length=0.05, angle=90, code=3)

## Warning in arrows(t * t, alpha - stdev, t * t, alpha + stdev, length = 0.05, : zero-length
arrow is of indeterminate angle and so skipped
```

```
plot(t*t, alpha, ylim=range(c(alpha-stdev, alpha+stdev)))
arrows(t*t, alpha-stdev, t*t, alpha+stdev, length=0.05, angle=90, code=3)

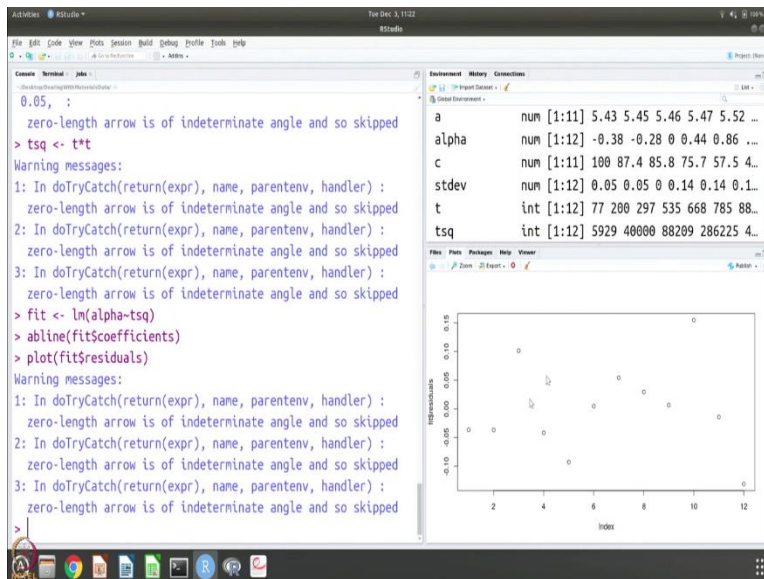
## Warning in arrows(t * t, alpha - stdev, t * t, alpha + stdev, length = 0.05, : zero-length
arrow is of indeterminate angle and so skipped

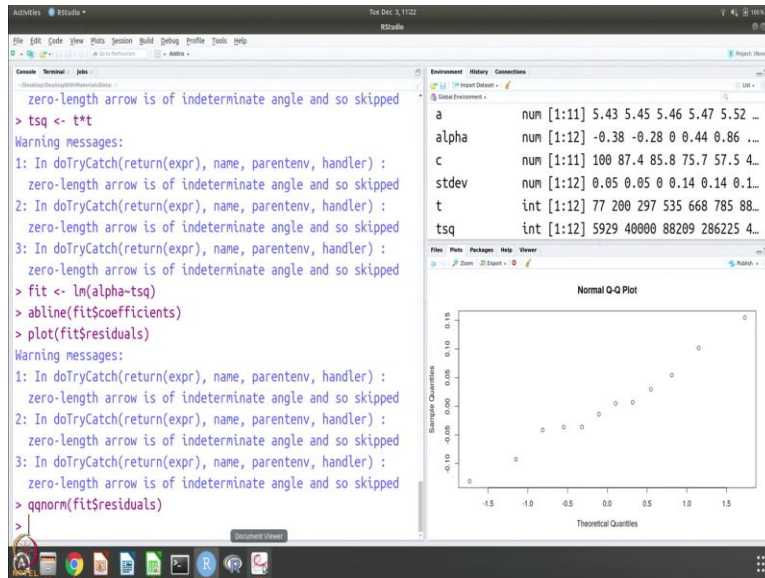
tsq <- t*t
fit <- lm(alpha~tsq)
abline(fit$coefficients)
plot(fit$residuals)
qqnorm(fit$residuals)
```



So, we have the data and now let us do the plotting and so I am going to take t square and I am going to fit alpha versus t square and this is how I have plotted the data, you can already see that this is becoming more or less a straight line, so it will fit a nice straight line.

(Refer Slide Time: 16:24)

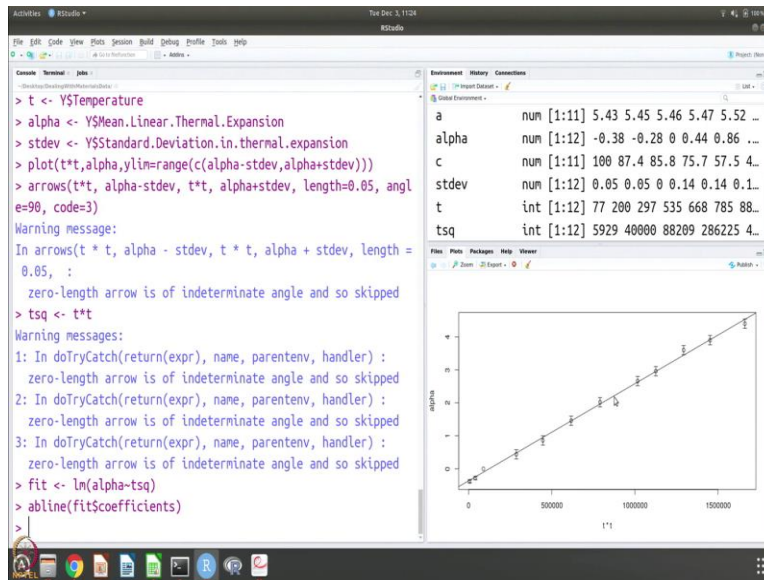




And of course you can look at the fit residuals and confirm that they are, you can see that about 0 on either side, they are there and they are randomly scattered and you can also look at the so this is like a straight line. Again indicating that we have got a better fit, so the purpose of this exercise is that whenever you get data, you should always plot it at see and most of the times by i you can do a very good fit. In the first case it was linear, so it was not surprising that we thought it is the linear fit.

In the second case, one could clearly discerned that it cannot be fit for a single straight line, then there are several different ways of fitting and we saw two, one is to fit for two different straight lines, so that you will get most of the data follow the trend. The other one is to see if you can make a higher-order fit. And in that case, we just took the square of the temperature and we saw that actually it does follow a nice straight line is that is the case.

(Refer Slide Time: 17:53)



So, and you can do the analysis after doing the fit and it is always a good idea to plot the data along with the fit to see that it is nice and this again gives you an idea that the errors are all random and it passes through closest to most of the points, so this might be a very good fit. So, in this case, we found that it is actually parabola that fits it better.

So, in many cases in material science and engineering you might already know what is the expected trend is, for example, sometimes it is Arrhenius type, so we know that if you take one by temperature and logarithm of one quantity, it should show a straight line and so on. So, in cases where it is not a straight line, there are sometimes methods to convert it into a straight line and even if it is some generic power law, logarithm always makes it into a straight line for example.

So, there are also other ways of exploring data, so one should always plot and see and if you do not know the, if you know the functional relationship you can try to exactly plot that and see. And if you do not know then you can explore and by looking at the plot most of the times you will be able to get the fit correctly or estimate the fit correctly and or at least if you make a fit and if you plot it and see you will know whether it is good or not just by eye. And then there are other ways of checking that your fit is good by doing the analysis on the data, this is just a starting.

So, we will continue doing more of the fitting and linear regression and analyzing the regression results and so on and so forth. But this is just a starting point of some linear fitting and some non-linear fitting, quadratic fitting, parabolic fit that we saw. And we also see that somebody might as

well fit it as two different straight lines for two different regimes and the data fits. So, one cannot say whether one fit is better than the other, except if you have physical reasons to believe that this is the right fit or if there are changes in mechanism that you do expect a linear fit in different regions and slope change between these 2 regions.

So, these are things sometimes known from your knowledge of material science and engineering and if you do not know, then you can do exploratory in which case you can try different things and whatever serves your purpose you can pick that. So, we will continue doing more fitting in these sessions to come in this module. Thank you.