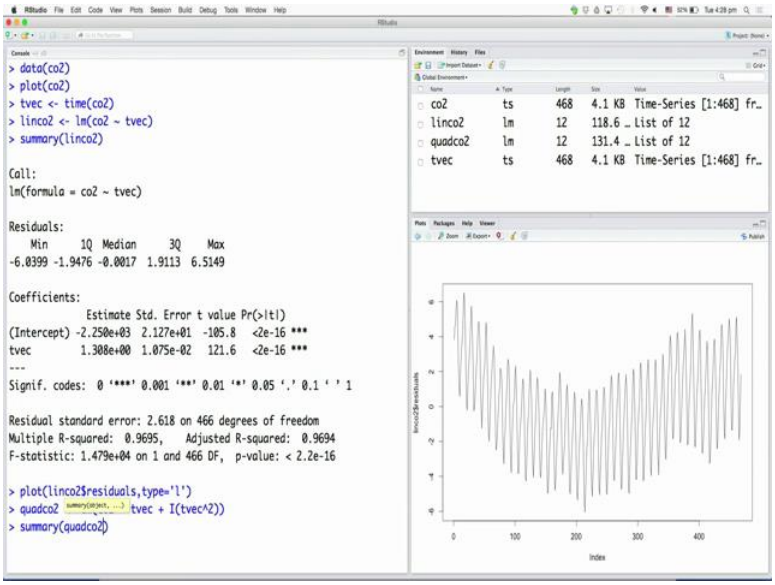
Introduction to Research
Prof. Arun K. Tangirala
Department of Metallurgical and Materials Engineering
Indian Institute of Technology, Madras

Lecture – 17 (3 B)
Modelling Skills

(Refer Slide Time: 00:40)



Welcome to the R session that supplements the lecture on Modelling Skills. In this session, we will work through two examples; in the first example, we will take up the CO 2 data set that we briefly looked at in the data analysis session. Once again, we start with a clean slate from R and load the CO 2 data set that comes with the base package. Always remember the CO 2 data set need not be just with one package, they are other packages when you load, you can also have the same data set. I am referring to the CO 2 data set that comes with the base package called data sets. So, it has loaded the CO 2 data. And to conform that's indeed the same data that we worked with last time. Just look at the plot and you must quickly notice that it is indeed the same one. It has a linear trend, atleast prima facie, but it could be a quadratic trend or a cubic one, you do not know, but definitely it has a trend function of time. On top of it, we also have an oscillatory trend which we would like to model. And then, there could be a stochastic component to it which also has to be modelled. So, the goal here is to model the series,

so that I can use that model for forecasting the CO 2 levels.

We shall not go through the complete modelling exercise, because a complete modelling exercise may also call for some theory of random processes, where we learn how to model the stochastic components. I will only show you how to model the trend and the oscillatory part, and then the rest is reserved for some other course or may be if you are already familiar with it, you can go ahead and do it.

So, here the first thing that we would do is we would model the trend, where the trend is a function of time. So, the time is a regressor and the CO 2 is the variable of interest. Or to extract the time vector we could use the time command. As you must notice here, in the work space panel, CO 2 is a time series object, and therefore, its always going to have time stamps as one of the attributes. We extract the time stamps and collect them in the tvec – it's just a variable name that I have chosen. Now, we could extract the trend using what a routine known as lm in R, lm stands for linear modelling or linear models. The estimation algorithm underneath this routine is a least square - standard least squares algorithm - with a lot of other features as well. You could, for example, model only using a subset of data, you could supply weights and so on, but we are going to use the most plain or vanilla version of lm. And let's call this model as lin CO 2 – that's just a variable name that I am choosing.

And let me show you what I am doing here. So, I say here I supply the formula that I want. Now by formula what we mean is the symbolic relationship between the predicted variable which is CO 2 and the regressor which is tvec. I do not have to tell lm that there is an intercept term as well. So, it is understood implicitly that there is an intercept term. On the other hand, if I do not want to the intercept term - for some reason I believe there is no intercept term - then I could use this syntax, but at the moment we do not know if there is an intercept or not; of course, we can look at things visually, but let us rely on the algorithm first to tell us if there is an intercept.

So, this is the model now we have. Now, we can examine this model and we should do it to proceed further. The summary command is a very multipurpose command which applies to different types of objects; lin CO 2 is an lm type object, basically it's a model

type object and it has several components to it. But first let's look at the summary and see what it is brings out. So, on the top, it gives you the formula which essentially is the symbolic relationship that we are modelling. And then, it gives you some statistics on residuals, what is the median and so on. You see some idea of whether the residuals are of zero mean, what is the range and so on, but we will come back to that. The primary interest for us is the estimates of the intercept and the slope. We have just fit a linear model.

And as you can see here, there are four columns under the coefficients heading. The first column gives us the estimates; the second column gives us the standard error as we call one sigma error or the average error in the respective estimates. And then, the last two pertain to the statistics on these estimates and there are also stars here. So, let's quickly understand what all of this is about. So, the estimates, of course, we can read off; by themselves the estimates do not carry a lot of information, they have to be interpreted in the context of the standard error. So, for what I mean by that is, suppose the estimate here was a very small value, you cannot come to the conclusion that the estimate can be neglected. Suppose slope turned out to be 10 power minus 4; it has certain units; so, it is going to be sensitive to the units. So, the value of something of the order of 10 power minus 4 by itself does not make any meaning or carry any significant meaning to it, unless it is interpreted in the context of the standard error.

So, standard error here is much smaller relative to the estimates themselves. Conveying the fact that these estimates are to be treated as significant, which means you cannot really treat them to be negligible or theoretically zero. Now the t value column is something that I will not go over, it's pertaining to hypothesis testing. So, may be in the course on hypothesis testing, you will find a lot of details on this. The last column, reports what is known as the p value; and the p value is again the probability of estimating this parameter, finding, obtaining an estimate larger than what we have observed, but without going too much into the details, the simple interpretation is, if the p value is extremely low, then the null hypothesis that the parameters are truly zero value. What are the parameters here? The intercept and slope, and that the entire purpose of this analysis is to test the hypothesis - that the individual parameters are truly zero.

So, when the p value is extremely low, the null hypothesis that the parameters should be zero value must be rejected. And if you recall, we use a phrase - if the p value is low the null hypothesis must go. So, it is just a catchy phrase to remember. And the three stars here are kind of telling you that the parameter estimates are significant. They all telling you the same thing; you can look at it from a hypothesis testing view point or you can look at it from a significance test for the parameters; either way what it's trying to tell us we cannot ignore the parameters estimates practically. Therefore, this linear model should have both slow and intercept.

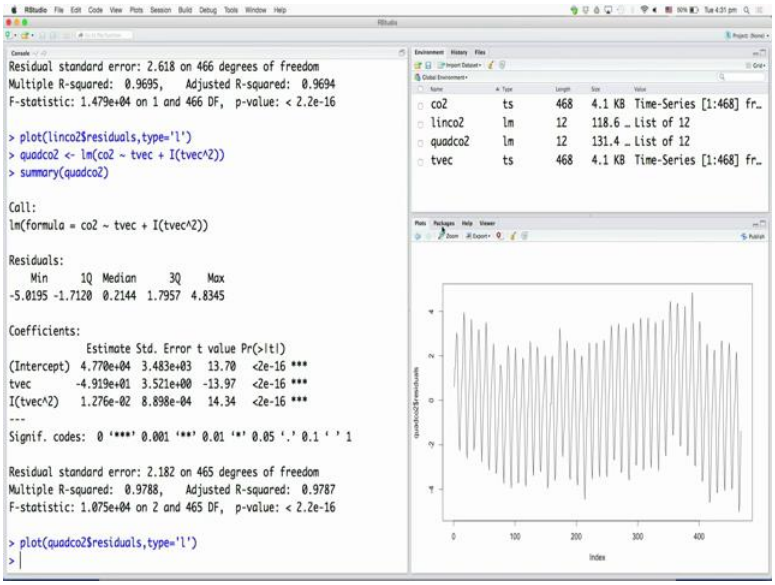
Now the question is - if this linear model is sufficient? We have discovered that this linear model as significant parameters on it. Now, to do that, we can look at the residuals; for example, we could plot the residuals; the nice thing about R studio is it tells you whether the field that I am typing is a valid field in the lin CO 2 model. And by default, it may plot a scatter plot. So, we will plot a line plot. So, you can see here the linear trend as been taken off, but it seems to be some quadratic trends as well, because a pure sine wave of that frequency cannot have a shape like that; may be we can fit a quadratic model as well, and check if the residuals are much better behaved and so on. So, to build a quadratic model, you can go through this syntax here, and this is something that you may want to pay closer attention to. Again, the same story; we use lm, specify the relationship between the predicted variable and the regressor; the regressor is as usual our time vector. This time around we want to built a quadratic model.

So, our model is y is some a, a 0 plus a 1 t plus a 2 t square. a 0 is implicitly understood to be present, so we present, we are supplying tvec here in the formula forcing lm to fit an a 1, and then we used this syntax I of tvec square. This essentially tells R that tvec square is another regressor; if we do not do this, and instead we say tvec plus tvec square what R would do is it would add up tvec plus tvec square, and treat that as a single regressor, and that's not what we want alright.

So, always remember, when you are creating regressors out of a single explanatory variable - here the time vector - then we should follow the syntax. Tomorrow you may have tvec plus may be cubic term or a logarithm term something like that; all of that has to be - each regressor - has to be encased in a similar fashion like this alright. So, we

have a model.

(Refer Slide Time: 10:54)



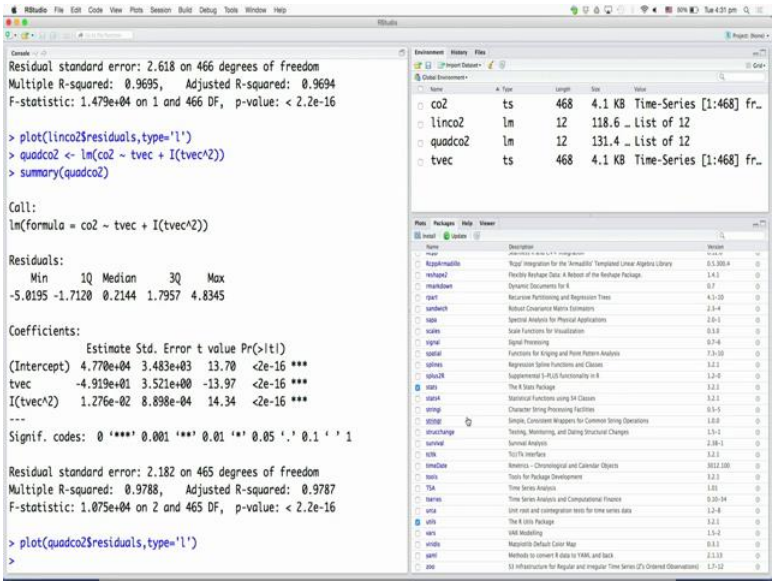
And now, we can once again look at the quadratic model, and look at the parameter estimates. Of course, technically speaking, we should look at the residuals first, and then the parameter estimates, but we are taking a slight deviation from that. So, again, here the stars for me indicate or the p values as well indicate that the parameter estimates are significant. That is, we can reject the null hypothesis that each of the terms in my model a 0 intercept, a 1 - the first coefficient corresponding to t; a 2 - the second coefficient corresponding t square, are all not zero or the estimates themselves are significant. We do not have, unfortunately, in this case a true model. I do not know or we do not know how this CO 2 was generated. The actual process is far more complicated perhaps then the models that we are trying to fit; remember that. So, there is no truth to compare and that's why we go through this hypothesis testing; of course, if we know the truth then all this modelling exercise is futile alright.

So, there are a bunch of other pieces of information here that are reported by summary, but to go over that requires a good set of lectures on linear regression, and therefore, I am avoiding that. But if you are already familiar, you will enjoy these pieces of information and make more meaning out of this analysis.

So, once again let's look at the residuals and see if this time the trend - the quadratic trend - has vanished right. So, this time, we look at the residuals; yeah the trend has vanished. There could be, of course, a cubic one, we could do that, and then go on and check if there was a fourth order polynomial trend and so on. I leave that to you; now that I will shown you how to do this.

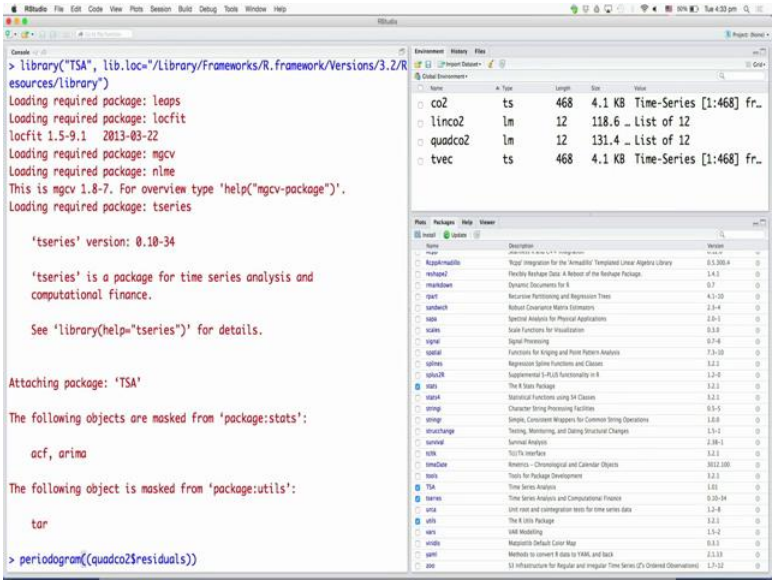
Let me move on to the next step where we analyse the residuals. For now, we shall assume that the quadratic trend is the only trend, but you should not do that, go further and fit a cubic trend. We will assume that the quadratic trend is only one and analyse the residuals. Now, when we look at these residuals here, we can clearly see an oscillatory behavior, which means we can now try to extract the frequency of this oscillation; that means, the CO 2 level has a periodicity to it. It repeats itself after a certain time. Now, how do we extract the frequency? I can do this visually, but that's going to be too rudimentary an analysis. Let's instead use the Fourier Transform rule. I don't know how many you are familiar, but if you are not, then Fourier Transforms allow us to detect the periodicity in a simple way by constructing what is known as a power spectrum or a power spectral density, where you analyse the contributions of different frequency components within the signal to the total power of the signal. And the simple rule of thumb, which has a theoretical basis to it, is in a power spectral plot if I see a peak at a certain frequency then that frequency component is significantly present in the signal.

(Refer Slide Time: 14:29)



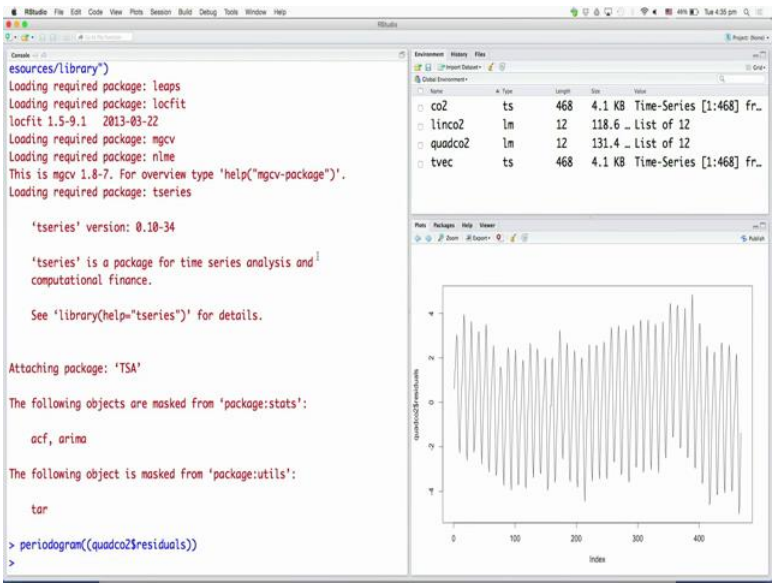
Now the base package in R does come with tools to perform spectral analysis or Fourier analysis; there's a command called spectrum, but I prefer the Time Series Analysis package and it's called the TSA. And I am going to load that, because it has a nicely wrapper or a nicely wrapped routine which is built on the base package to do this spectral analysis.

(Refer Slide Time: 14:43)



So, I have loaded the TSA package, and you should be… you should see this kind of a display. Now, when I load a new package in R, always there is a possibility that some of the routines in the new package share the same name with the routines in the existing or the base package. And R gives you a warning that the following objects are now marked from this package and so on. So, basically, they are telling you that acf, arima, tar and so on, if they are of interest to you, there is an overlapping between these two packages. So, you have to be careful and so on, but, anyway, we will not go into that. But one thing that I want to tell you, the TSA package also has a CO 2 data set, which looks quite a bit different from the CO 2 data that we have been working with.

(Refer Slide Time: 16:02)



Anyway, so, let's quickly look at the periodogram, the command is periodogram. And I am going to really directly pull the residuals here, periodogram of quad CO 2 dollar residual; you can see a dollar operator is being used to extract the component. And beautifully it shows me the frequency content of the signal. So, the way to interpret this plot here is on the x-axis I have power, and on the… sorry, on the y-axis I have power, and the x-axis I have frequency. And what it's basically telling me is that at this frequency, roughly, I have the maximum power; and this obviously, there is don't need to guess, is the main frequency component that we are interested in. But there is something else also this spectral analysis reveals which is a presence of a harmonic

alright. And this is, perhaps, well to conform whether there is harmonic or not - harmonics are integer multiples of fundamentals.

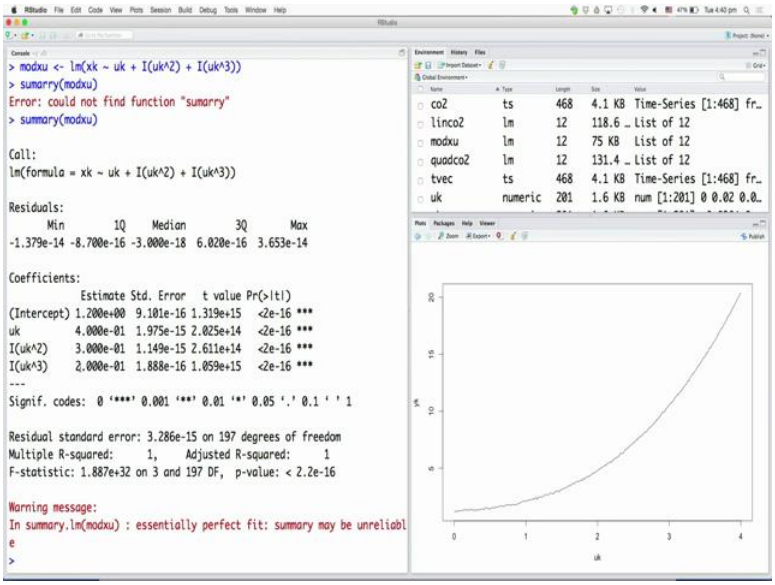
So, if you think of this big one - the frequency at which there is a big peak as a fundamental - then approximately the tinier one to the right of that, can be thought of as a harmonic. And typically, presence of harmonics means some kind of non-linearities in the generating phenomenon and so on, but it's very interesting, because visually we could not see this peak. So, you see, mathematical analysis brings out what we cannot really visually see, but at the same time visual examination is very important.

Now, if you look at the power contributions from very low frequencies - almost near zero frequencies - there seems to be some significant contribution that is perhaps indicative of a very low frequency trend, which probably is indicating that we should have also modelled the cubic one; even our visual analysis, in fact, reveal that if you go back to the plot, there seems to be some kind of a cubic trend. In fact, if you plot the cubic one, I should tell you that you will find the a three - that is a third coefficient, the forth coefficient - also being significant.

So, try it out, try fitting the cubic one and the fourth order one, and see what the analysis tells you whether the trend is indeed a third order or a fourth order polynomial, and then we have already learnt how to analyse the residuals, extract. So, now, you use these models to remove the trend and the sinusoid, and then you can see the residuals, which of course, I won't go through at this moment, but you can always write to us if you want know how to do it in more detail. We can send you the script.

Alright then, so, let's move on to the second example. In the first example, we have learnt, essentially, how to perform linear regression in a systematic manner in R. We are also going to do one more linear regression here, but this example pertains to that of over fitting example that we went through in the lecture.

(Refer Slide Time: 18:53)



So, let's clear the screen here, and generate the data required for the over fitting example. What I am going to do is, I am going to generate the input as a sequence here, from 0 to 4 in steps of 0.02. So, I have 201 values of uk, and correspondingly I will generate the noise free response. If you recall, go and refer to the slide which gives you the relationship or the data generating expression for the noise free part, and this is what we had, and then let us call that as xk. Now, as we have always said, we never get access to the true response, we only have access to the measured value of it. So, to make the situation realistic, now we add noise. And we add noise in this fashion right; we add Gaussian distributed noise of the same length as xk. And we adjust the amplitude of the noise such that the signal to noise ratio is fairly high, and that's what I am doing here.

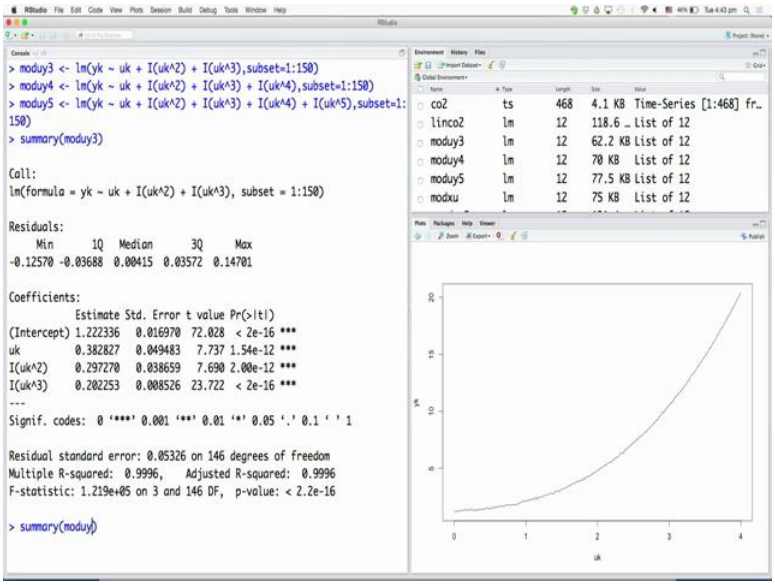
By default, if I don't have the scaling factor there, the variance of vk would be one; it's a unit variance random noise that I have been generating. So, let us generate the measurement; here I have yk equals xk plus vk and that's the noise part I have, noise or the measured value. So, let's simply plot here uk and yk, and see if this is what indeed we wanted yeah. So, it looks pretty similar to what we we wanted; of course, it looks a bit noisy, not so noisy. You can add… if you want, you can increase the levels of noise and repeat whatever we are going to do.

Now, let me go through two or three different models and then will conclude the session.

So, the first model I just want to show you, if you had access to the noise free part, and you had and you knew that the relationship is a third order polynomial, then you can… let's do that here okay. Let's call this as mod xu, and call the lm, and supply the symbolic relationship. So, we have here xk tilda uk plus I uk square plus I of uk cube alright; so that's it alright. So, that brings up the summary for the model that we just fit between x and u. And I quickly want to draw your attention to the warning message at the bottom of this display. It says that essentially a perfect fit. It has to be, so what it says is, the relationship is perfect, there is nothing left unexplained, there is no residual from this model so to speak, and that's also reflected in the median of the residuals are extremely small. There is zero up to the numerical precision. And also, the parameters estimates have been identified in a perfect manner and so on, but this is all a utopian world, we don't have access to the noise free path. So, we move now to reality, am just showing you, that if you had everything in hand, the lm will perfectly identified that for you.

So, let's move on to the over fitting demonstration or the over fitting example. In the lecture, I pointed out that when we over fit, that is, in this example if I fit a fourth order polynomial or a fifth order polynomial I am over fitting; if I fit a second order or a first order, I am under fitting. So, in practice - do we know what is a order of the polynomial? No, we do not. So, how do we go about determining it. Let's assume that we are given there is a polynomial relationship, even if you are not given, the plot of y versus u tells me or strongly indicates the polynomial kind of relationship. Now, I start guessing; we necessarily, obviously, rule out the first order, that is a linear one. We can begin with the second order, third order, fourth, and fifth, but I will skip the second, because the primary purpose is to show over fitting not under fitting okay.

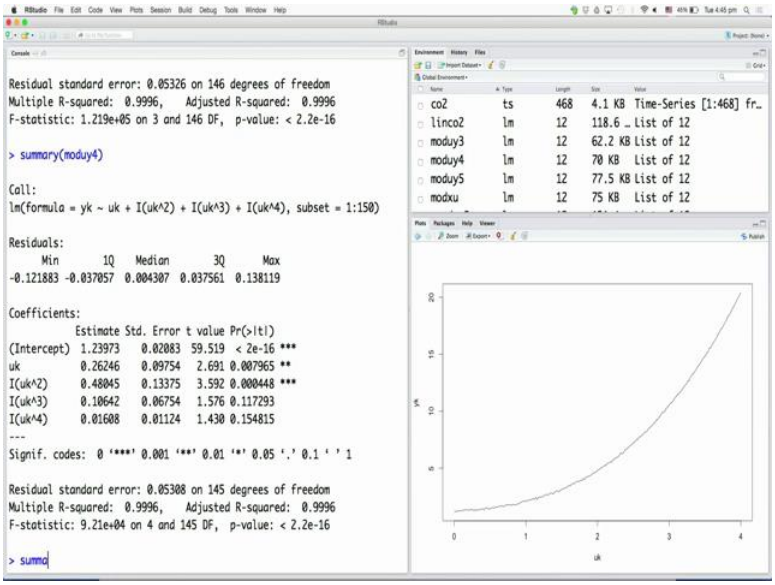
(Refer Slide Time: 22:35)



We will first identify the cubic polynomial relationship between y and u, and then build the other two models, and let me do that here. mod yu 3. So, I have here. And also, what I am going to do, is I am going to pick the first hundred and fifty points for building the models and use the remaining fifty-one for prediction. This is something that's probably going to be useful for you. And that exercise is also necessary, as I have mentioned the lecture, for cross validation. So, in lm, I can exercise this option by saying subset use a first one fifty points; you can use any one fifty points or even change the length and so on. Okay so, I have this mod uy 3; let's also build the other models right here in a similar fashion and the fifth order polynomial as well. So, I am going to build all of them at once. So, these are the three different polynomial models.

Now, we want to ask how good these models are right. A good test for these models, of course, is to look at the individual model - the parameters estimates in these individual models. So, here is a summary of the third order polynomial fit, and you can see that the parameter estimates are close to the truth. In this case, I know the truth because I have simulated the data and the estimates are quite close to the true values; you can never expect them to be exactly equal for obvious reasons. And the stars here also indicate along which are coming out of the low p values that the estimates are significant, good. So, third order polynomial is good.

(Refer Slide Time: 25:17)

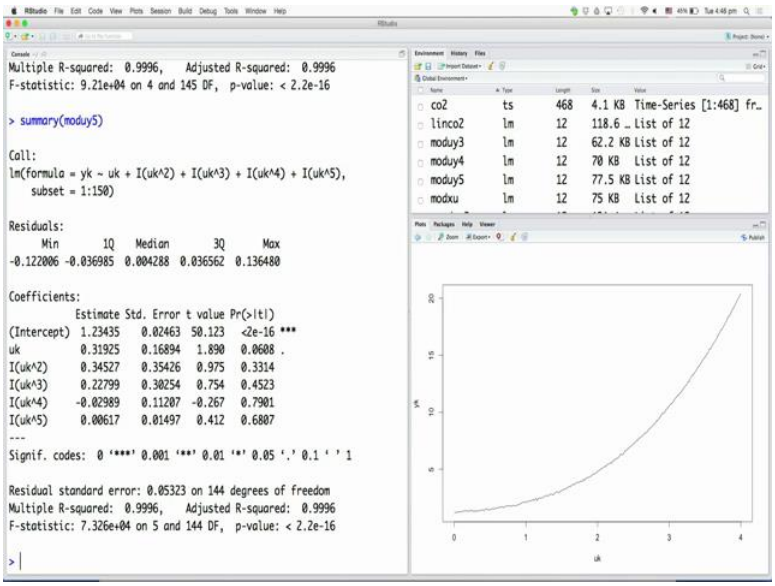


We can also look at in a similar way the fourth order. Now, you see something different. I have over fit, because this is the fourth order polynomial look at what is happen. The intercept term is pretty close to the truth, and then starts the jumble tumble here, that we have the estimates deviating from the truth significantly to the extent that the coefficient on the third order term is not significant; you can look at the error. The error in the estimate of the coefficient for u cube is quite significant compared to the estimate itself; and like wise for the fourth order coefficient as well. And that's why there are no stars here telling you that the null hypothesis, that the coefficients on u cube and u to the power four are zero, but we know that that is partially correct, but atleast we know for sure the coefficient on u to the four should be zero. What we should take home from this simple exercise is, when we over fit, we can have a situation where the parameter estimates, even though the true parameters are not zero, can turn out to be insignificant.

So, we do not know how this over fitting is going to hurt, and I always like to give this example, that data is the food for identification and parameters are guests. So, when you do not have enough food for your guests, then some guests may go satisfied and some others may go hungry. We do not know who will go hungry and who will go satisfied, but the bottom line is they will always be some guests who will go hungry and the guests here are parameters. What we mean by hungry is here large errors; we do not want that;

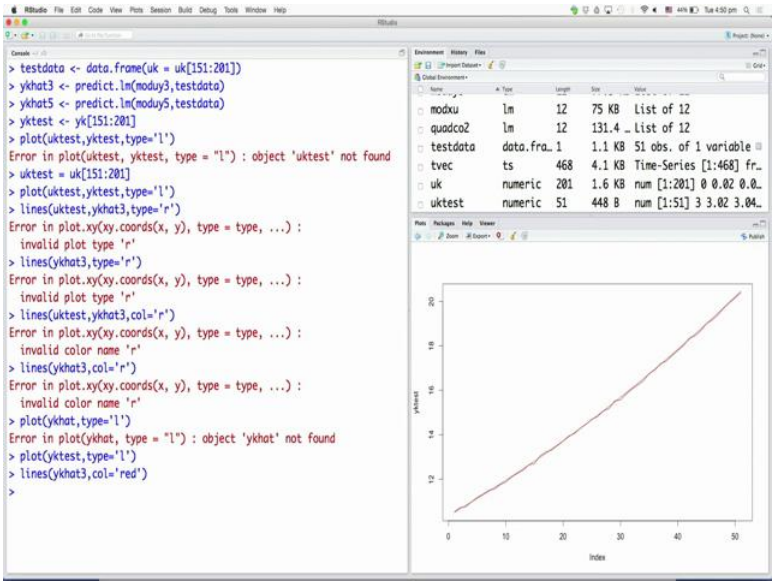
when we invite guests, we want all guests to actually go satisfied. So, always remember, data is food for identification; we don't want to over invite, we may under invite, but that is also not good in modelling. We want to invite exactly the number of guests that are meant for the food because you don't want any wastage of food either you have remember. And that analogy, hopefully, will stay with you whenever you are modelling.

(Refer Slide Time: 27:47)



And we should expect a similar behavior for the other one as well. So, for the fifth order polynomial as well, you can see none of the parameter estimates except the intercept term are significant, which means they have large errors in them. So, this has, obviously, gone for a full toss; we can compare, I will just compare the predictions. I will show you how to compute the predictions of the models. And we can quickly compare the predictions of the third order and fifth order polynomial on the remaining fifty-one points or we can choose the full data sets. We will just choose the remaining fifty-one points.

(Refer Slide Time: 28:25)

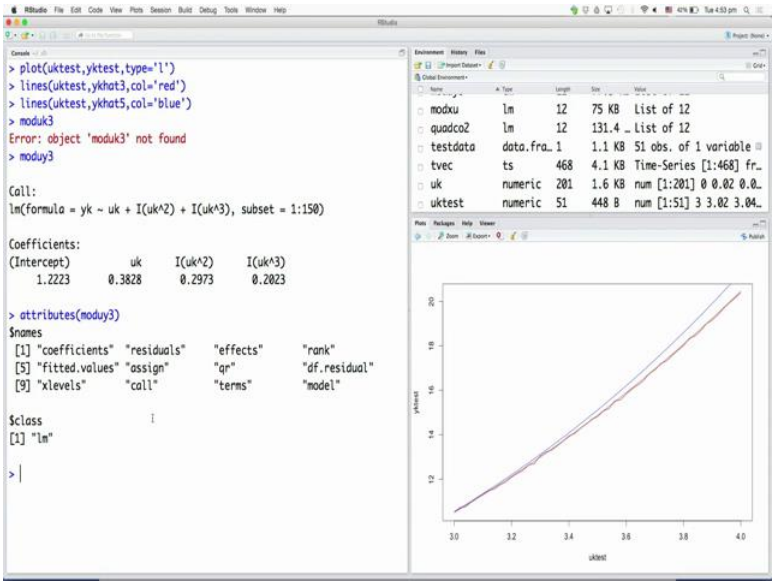


Let me do that. So, we call this as yk hat. So, let me... first the procedure to compute a prediction is to first create a data frame. Remember, that R accepts data frames for modelling and predictions and so on. So, lets create a data frame, where… I can simply say here uk 151 to 201. So, those are my… that's my test data uk. And now, I can make the prediction here alright. So, mod uy three dot test data predict dot lm or simply predict; predict knows by looking at the object that you feed, which is moduy 3, that's a model; it understands that it has to work on a linear model object. So, predict dot lm explicitly is telling R that the object or the model that I am supplying is a linear model type, and then, I am supplying the test data on which it has to compute predictions. So, I have done that. And similarly, let's do this for the fifth order as well. I leave it you to compare the fourth order; something to work on for yourself.

Okay so, we call this as yk hat, hat is typically used in estimation for a prediction approximation and so on, and I have used the same convention. Now, let's compare the predictions here. So, I have a plot. Let's plot the yk first. In fact, let us call also yk test; create a variable which corresponds to the yk for the test data alright. So, now plot uk test or yk and test and; this is not uk test; we have not created, sorry. So, here is the plot for the 151 to 201 points between yk and uk. This is a true, the measurements. Let's ask what are the predictions of the third order and fifth order polynomial models. We can use

here lines and say uk test, yk hat 3, and we can use a red color to see how the plot is. So sorry; this is color. So, let's take a different approach here. Let's plot simply the yk hat, yk test, and then plot yk hat 3. So, we can compare the predictions here. I have plotted yk as the function of time.

(Refer Slide Time: 32:08)



But we might as well plot as we did earlier here. We can plot this as a function of u as well okay. And then, here lines. Alright great ! And then we can also plot, maybe, on blue - in color blue; the prediction from the fifth order polynomial and see where it has gone. So, you can see, that in this case the predictions of the fifth order has gone apart. In the lecture, we have seen that the predictions had gone unstable as well. Now, how these predictions behave completely depend on the realization of the noise that we add, but the fact is, first of all, the parameter estimates are not reliable in the fifth order case. And therefore, we should not even trust the model; even if you were to trust that the predictions are not so great. I can tell you, if you were to repeat this exercise, these plots here will look different and the estimates will look different, because you will generate a different realization of noise.

But the bottom line is, if you were to over fit, there are two symptoms that will be clearly visible. One, that you would see the large errors in parameter estimates, which means

you have over estimated, you have over parameterised your model; and two that the predictions will fail somewhere between miserably to not so miserably, but they will be poor on a fresh data set. Always, in modelling therefore, if where ever possible have a cross validation data set; of course, some times may have two smaller data, but it should always be a practice to have a test data set, where you can test the trained model on the test data pretty much like we test students, trained on assignments, on exam papers right.

So, go through this exercise by yourself and get a feel of what it means to model. I just also want to show you whether it is a lin CO 2 or the models that we have been fitting for example, mod y k 3 sorry uk uy 3. If you just type, this is what it would give you, but if you want the full list of components in mod uy 3, attributes will tell you what are all the other things that are contained in each of these models that are returned by lm. You have coefficients, the fitted values, you can compare the fits; the fitted values are the predictions of the model on the training data; you can extract the residuals like we did; then you can also get other details as to how it was called and so on, depending on how well versed you are with linear equations.

There is also a non-linear version of the lm routine and there is a generalized version called glm. The non-linear version would use the non-linear least squares method; hopefully, some day we will have a chance to go over those as well, but we will have to conclude.

And let me do that by saying, by hoping that you have enjoyed this modelling session and write back to us, as always, with any questions that you may have pertaining to modelling or how to fit models in R and we will definitely be glad to help you.

Have a good day.