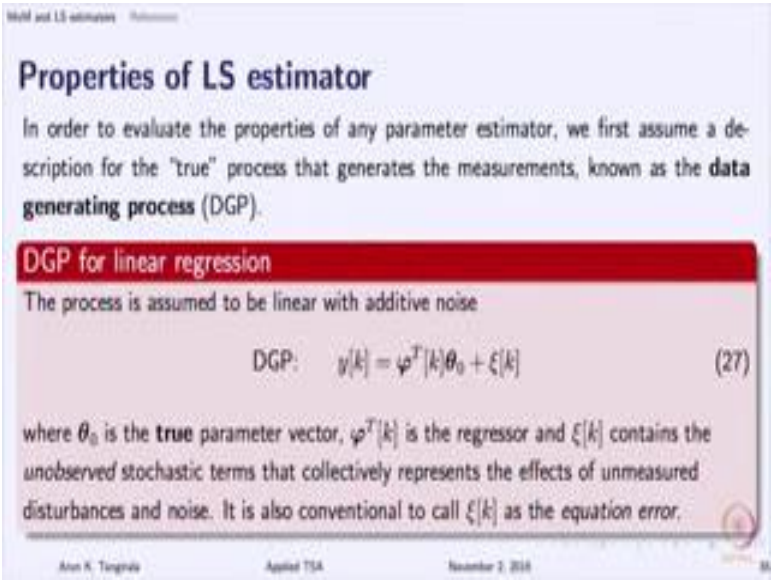
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 99
Lecture 43A - Estimation Methods 1 -6 (with R demonstrations)

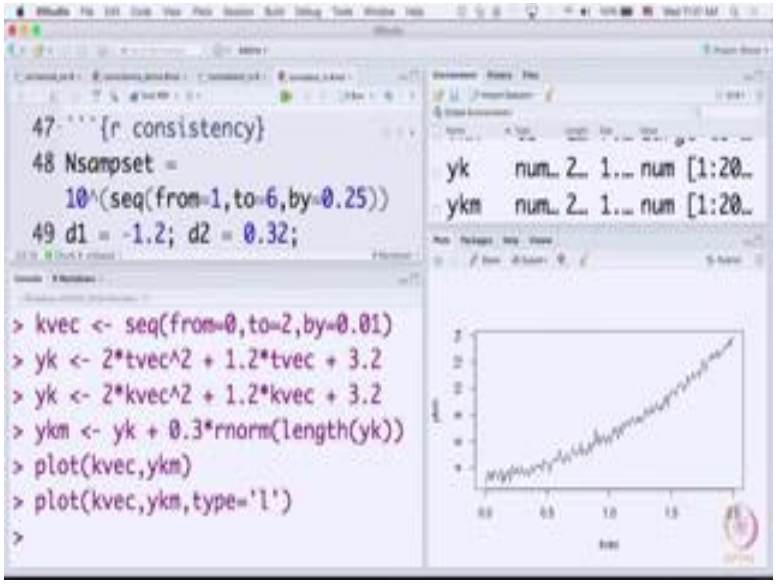
Good morning, what we will do today is we will discuss exclusively the properties of least squares estimator and then if time permits we will look at weighted least squares as well. So, what we discussed yesterday is of course, a solution to the OLS problem and also the goodness of fit measures, mainly the r square and adjusted r square. The r square tells us how good a fit, the model has obtained with the help of least squares; you see the model by itself cannot really understand the data; you need an estimator which essentially trains a model to understand the data.

(Refer Slide Time: 00:58)



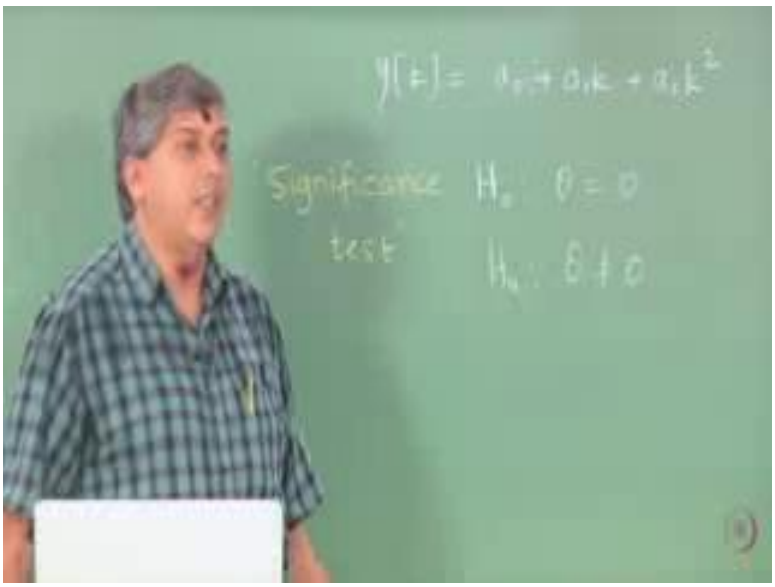
And now at the moment we are talking about the least squares estimator and r square measure and adjusted r square measure tell us how well the model as understood the data and the difference between r square and adjusted r square is the penalty term that is kind of included implicitly into the or explicitly I should say in the adjusted r square, whereas in the r square measure; you can achieve high r squared measures with over parameterized models and r square would not tell you, but it is happening.

(Refer Slide Time: 01:41)



So, let me actually quickly show you an example in r where let me here alright. So, here we are just going to work with a synthetic example where it is a very simple example.

(Refer Slide Time: 01:59)



We consider some series y which is being constructed as some a naught plus a 1 t plus a 2 t square let me say k here k square and what we will do is we will generate the time vector first. So, I am going to do this and let us say and call this as kvec and let us generate y k. So, I just arbitrary chosen the values of a naught a 1 and a 2 of course, it is a quadratic, let us assume that we do not have y k, but rather measurements of y k; where

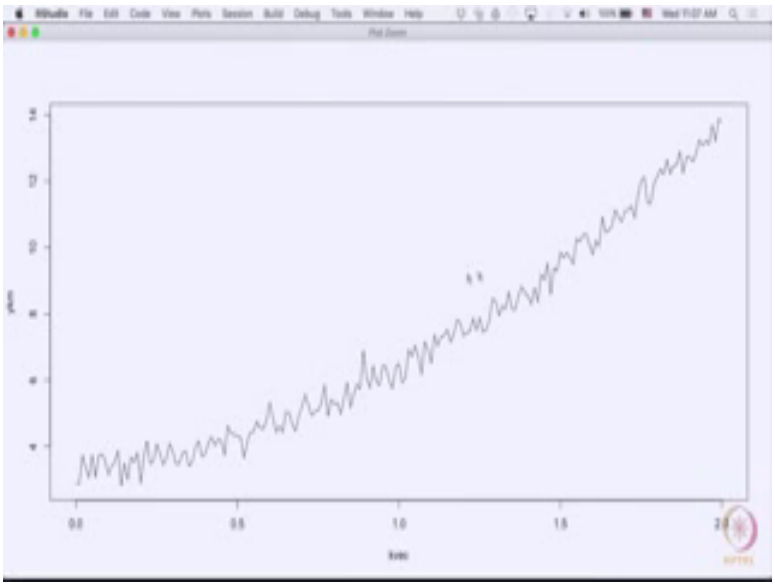
we add some noise to y k as you can see; the second term represents the you can think of it as measurement error or whatever cannot be explained by a quadratic polynomial.

You could adjust this factor here point 2 to any value that you want, but remember that as you are increasing that factor you are raising levels of noise and lowering the signal to noise ratio and therefore, as you increase levels of noise; you should expect poorer and poorer fits because our regressors are going to be 1 k and k square; those are the regressors that we are going to use yes.

Student: (Refer Time: 03:25).

I will change sorry, earlier I had t right thank you alright. So, let us generate here we just arbitrarily choose a 0.2 or even 0.3 and it is a good idea always to plot the obtained measurements as a function of the time and as you can see here may be it is better to plot a line type plot. So, as you can see here there is a trend right let me zoom this out for you zoom this in.

(Refer Slide Time: 04:03)

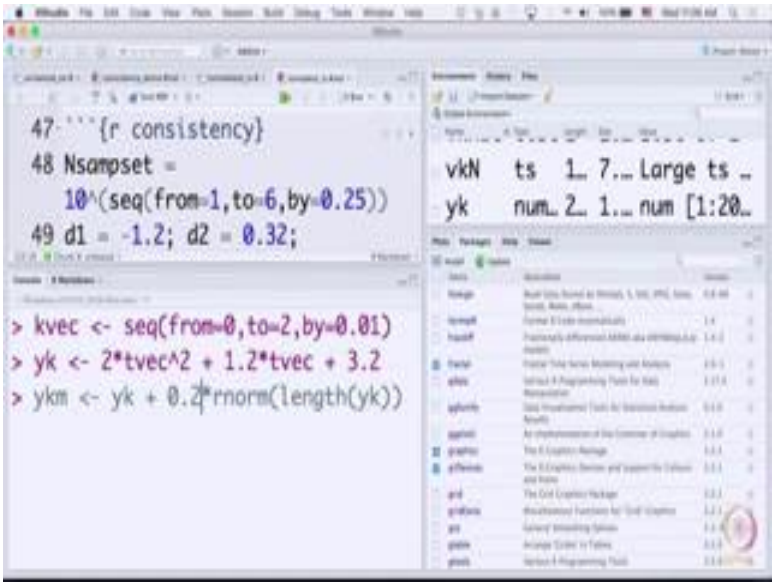


And then, so you can see that there is the trend here; there is a quadratic trend which is what we want the model to capture, but there are also local variations. Now under fitting happens when we miss out this quadratic trend may be we if it a straight line, over fitting occurs when we start modeling the local variations which are essentially due to noise and that is what I always say when we are working out assignment problem; there is a

concept involved underneath every problem and there are numbers specific to that problem. If you understand the concept then well and good; you have gotten the purposes serve, but if you try to now start memorizing the numbers then that essentially is over fitting.

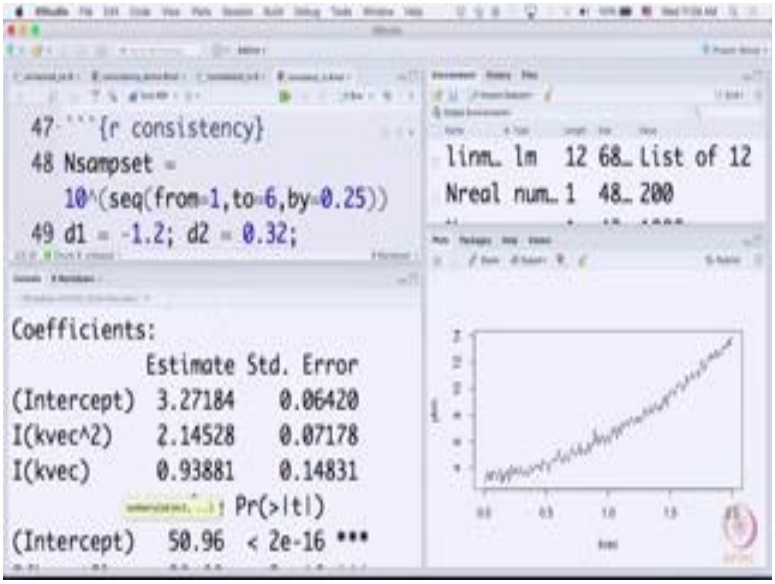
So, let us actually assume that I know it is a quadratic visual inception and then fit a model and as I have already shown you once how to fit linear models using least squares the routine is l m.

(Refer Slide Time: 05:07)



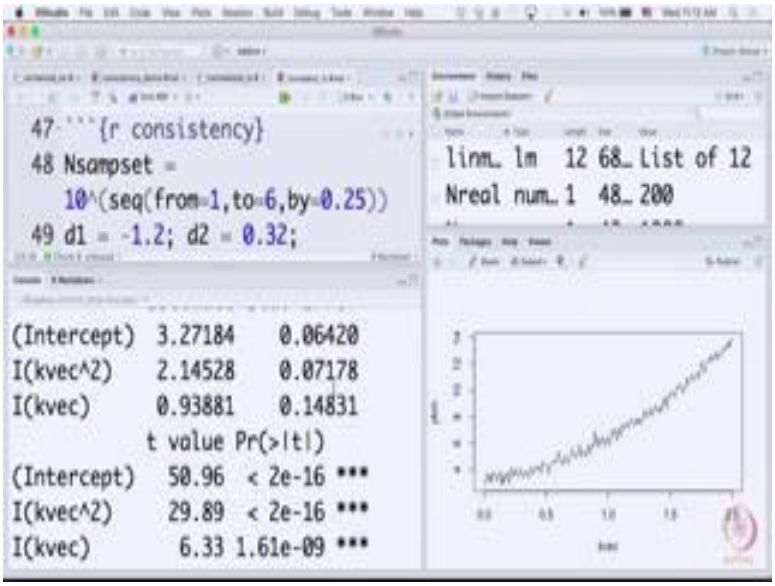
So, let us say here say lin mod y k, l m, we use the measurement and then we pass on the regressors in this fashion. Remember you have to tell l m that there are at least three regressors, but the third regressor need not be mentioned which is one that is there by default as an introspect term, so we have to supply the other 2 regressors and that is it. So, you would be making a mistake if you simply supply kvec square plus kvec because then it will club those two and think that as a single regressor alright.

(Refer Slide Time: 05:55)



Now, it is time to pull up the summary on this. So, when you do this the summary brings out a lot of interesting pieces of information relevant to your model. First it gives you the estimate essentially you are a naught hat a 1 hat and a 2 hat in the first column and then it also calculates the standard error. So, today I will talk about the theoretical expression that allows us to calculate this standard error. Remember we have what kind of noise have we added here; colored or white; we have added white noise. So, when we go to theory we will see that results are obtained only for white noise generally available for white noise type of errors.

(Refer Slide Time: 06:49)



When it comes to colored errors what happens is also said in the theory, but we do not have an expression for computing the standard errors because we will see that least squares estimates are efficient only when the errors are white. In other words, it gives you minimal variance estimates when the errors are white. So, first let us go through this quickly and then go back to the theoretical expression; we will also talk about by bias a bit shortly.

So, here you have the estimates and the standard errors remember as I said always in any model estimation exercise you have a hypothesis test underneath and that hypothesis test is concerned with essentially the parameters being 0 valued or not. Whenever you have a hypothesis test of this form whether it is multiple parameters or single parameter this is not only called hypothesis test, but also known as significance test because what you are asking is if the parameter estimate that has been obtained is significant. Even if the truth was that theta is 0 you would obtain some non zero value of theta right that we already know. So, question is whether the obtained value of the estimate is significant; the moment you hear the word significant you should think of this hypothesis test.

So, here we are asking if a naught is 0, a 1 is 0, a 2 is 0 we are conducting such kinds of tests either you can look at the standard errors and construct the confidence regions, but what is required for constructing the confidence regions the distributions of the estimates. Again theory will tell us very soon what is the distribution of the parameter estimates obtained using the least squares method and it turns out that fortunately it is just Gaussian distribution.

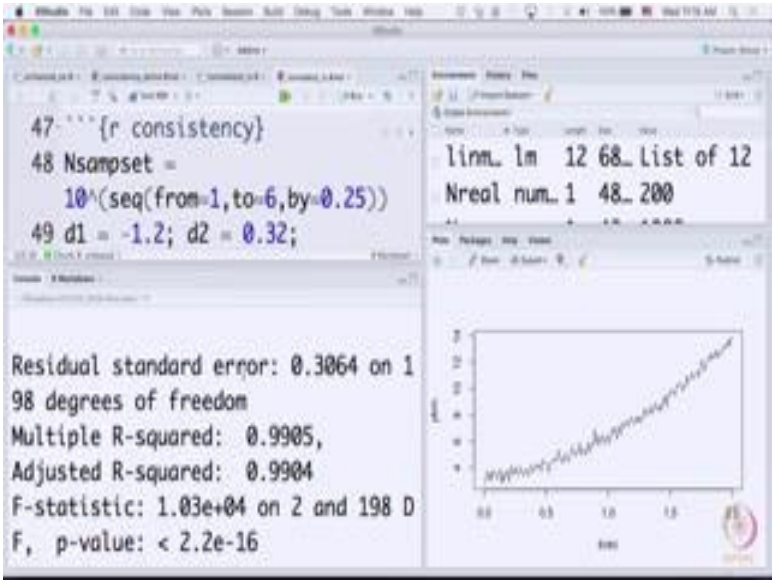
So, here for example, if I want to construct the 95 percent confidence interval; I can take the estimate obtained plus or minus if it is 95 percent roughly twice this standard error right 1.96 times sigma theta hat is what you want to add; this is sigma theta hat that you have. And quickly you can check if 0 is contained in the confidence region because what we are testing for in a hypothesis test, if a the postulated value is 0 and remember there are three different ways of performing the hypothesis test of which we are talking of the confidence interval approach. So if I construct the confidence internal here, what do you think for any parameter that is being estimated here would the confidence region contain a 0; 95 percent confidence region? No, so which means that essentially we reject the null hypothesis that a naught is 0, a 1 is 0, a 2 and 0 individually. So, all of them are significant; that means, I have to include those regressors in my model.

Very soon we will over parameterize that is we may fit for example, the forth order polynomial; instead of a quadratic and see if my significant stress tells me that I should have excluded those regressors, anyway. So, you can also look at the p value as I have mentioned this before long ago when we were discussing on stationary models and removing linear fits; at that time I had pointed out as well that you can look at the p value; if the p value is low then the null hypothesis must go right low meaning; low compared to your significance level; if you set your significance level to 0.05 then all the p values are lower than that which means the null hypothesis is rejected. So, again a different way of conducting your significance test here and the stars are telling you that they have already done that for you. So, the three stars would mean that these are like the star coefficients for you; they have you cannot really exclude them.

The t value that is being reported here is the t statistic, remember that is the third way of conducting hypothesis test, which is based on the critical value approach. You compute the statistic and compare it to the critical value. The statistic is being reported here, what is missing here is the critical value you can generate those critical values in r by turning to q t in this case, it although I said earlier the theta hat follows a Gaussian distribution when the variance of the corrupting noise is unknown theta hat follows a t distribution. We have talked about this in the context of sample mean, when we were conducting hypothesis test on sample mean; we had two situations variance unknown and variance known. When the variance is known sample mean follows a Gaussian distribution, when the variance is unknown standardized sample mean follows a t distribution.

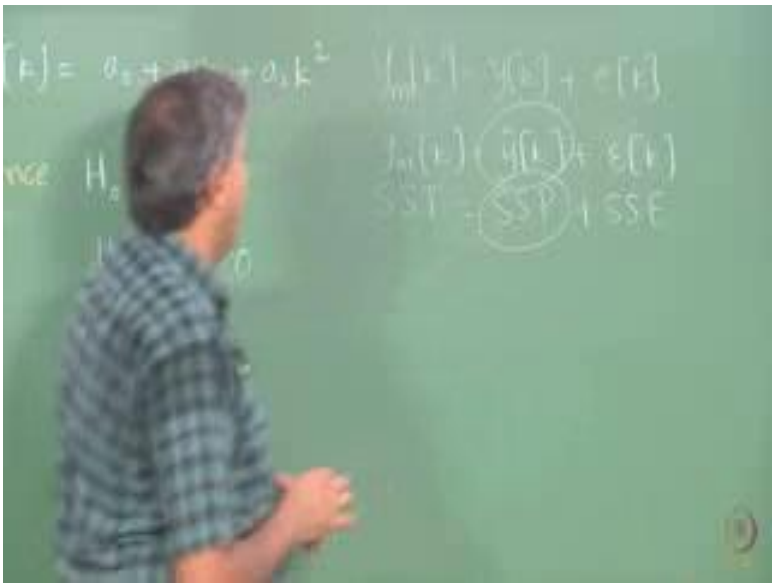
So, in general the statistic is calculated as a t value rather than z value, but when you have large number of observations then it does not matter, you can always refer to the critical values in the Gaussian distribution.

(Refer Slide Time: 12:06)



Anyway, so the significance codes are given and as you scroll down you are given some more pieces of information for example, you are given the residual standard error. What is this; what you think this is?

(Refer Slide Time: 12:25)



Remember that we are working with not y k, but we are working with measured y k which is y plus e k of a certain variance. Of course we assume it be 0 mean right; e k has certain variance sigma square e which I do not know; what is the variance that we had used. So, this is residual standard error which is 0.3064 on how many degrees of freedom

198 degrees of freedom. Now what is this is what we learn in the theory and then you will understand this value; I tell you shortly what is this residual standard error.

Then you have the multiple r square; this multiple r square is essentially is your r square extended to the case of multiple regressors. Here we have two, in fact if you exclude the intercept term you have two regressors. If you have a signal regressors then you talk of r square, again it is a measure of how good you have the model has, how well the model has explained the data and adjusted r square is a basically adjustment for the number of parameters. In this case they are pretty close because we are not over parameterized and finally, you have of course the adjusted r square that tells you that you have achieved an extremely good fit and we have managed to do that because the noise levels are low and we have included the correct number of regressors.

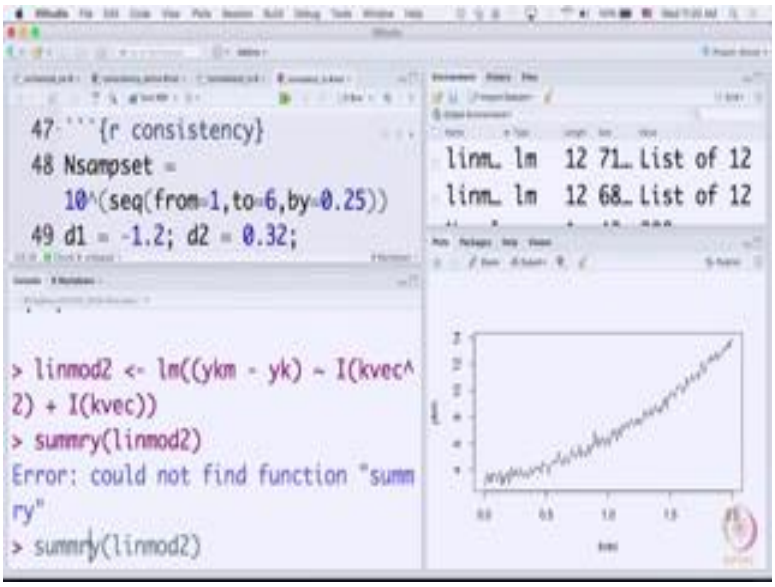
Finally you have this f statistic. This f statistic is another way of performing the significance test; f statistic earlier we did this significance test using the t statistic that is by directly looking at the parameters. We can also conduct this significance test or this kind of hypothesis test by asking how much variance, whether the variance explained by the regressors is more or less or is identical to the variance that is due to the randomness in the data. So, remember we did this earlier yesterday we wrote this expression sum square total is sum square predicted plus sum square error, when you are using the least squares method.

So, what we are doing here is we are asking if the regressors are just explaining the data by chance and if they are doing that then both these values should be theoretically identical because if this regressors I am sorry this is not the one. So, you are writing here y m k as y hat plus epsilon k, so sum square prediction is coming due to is the contribution due to y hat and sum square error is a contribution due to your prediction error.

The simple logic is that if the model; if the regressors that I have included should not have been included then the variation explained by the regressors should be more or less the same as the variation due to the errors because they just explaining by chance, but and other wise there is no systematic relationship; f statistic is a measure of the ratio of the variance explained by your model to the variance that has gone unexplained. If none of this regressors should have been included ideally that ratio should be 1 alright because

if there is no, suppose you were to fit a model directly on to some noise and you have included these regressors; try this out when you go back simply generate noise and regress noise on to your kvec square and kvec and see what the results tell you, it will tell that you should not have included those.

(Refer Slide Time: 16:40)



In fact, if you are confused let me just show that; so we have here let us say some linmod 2; we can say just generate some noise here or will take this specific noise that we have y k m minus y k being regressed on kvec square plus kvec. So what are we doing here, we are just regressing noise on to this regressors here it is a; obviously, it is a worthless business, but so there is some purpose to it in at least in terms of learning.

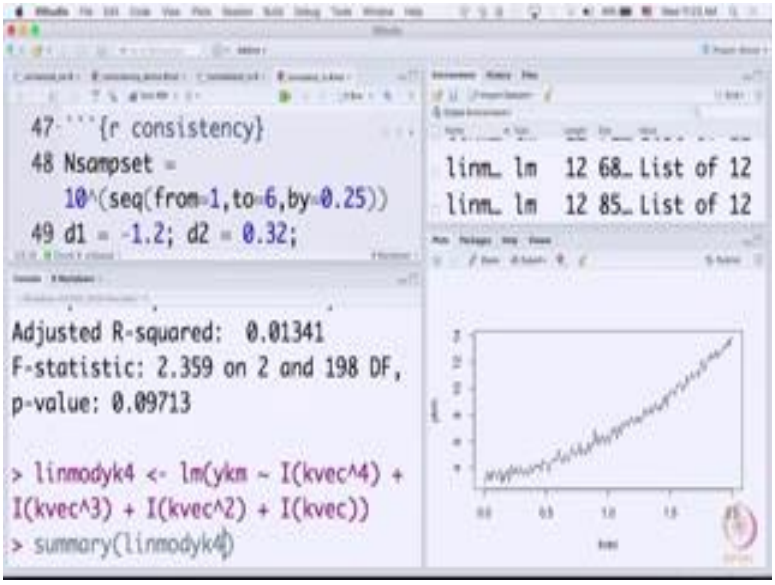
So, what do you see here right I mean kvec square just explains by chance, but if you look at the estimates and you look at the standard errors and so on, it says your intercept term should not be in a model, the regressor kvec is not significant because it is not star associated with it that should not be in a model, there is one star, but that is if you look at the legend here that is for you have here the legend for 95 percent; so you have 0 is triple star double star is for 0.001 and so on.

But all of them should work out correctly in your favor; in the sense if you go to the f statistic, if you look at the adjusted r square how much as these regressors manage to explain.

Student: (Refer Time: 18:20).

Pretty much you know insignificant that is telling you that this regressors have no role to play in explaining the noise, this is what would happen if you have included a completely wrong set of regressors for your model and the f statistic also says here, if you look at the p value associated with f statistic; it essentially tells you that you have not explained anything in your series, in this case it is pure noise with the help of your regressors. So, what I am trying to say show you here is, how to read the output of a linear regression. You have to look at estimates, you have to look at the adjusted r square, you have to look at f static and be convinced that the regressors are significant. Now let us turn to over fitting very quickly and see what kind of parameter estimates do we get when we end up over fitting.

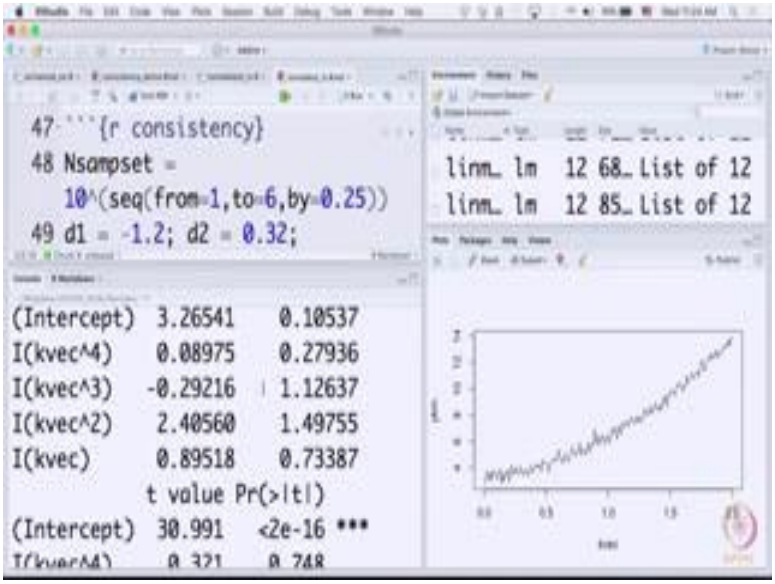
(Refer Slide Time: 19:21)



So, for this purpose; let us go back to our original model here and now include the power of four terms as well and let us call is as forth the forth order, alright. So, what do you expect to see now as far as the parameter estimates are concerned; the estimates corresponding to the fourth and third order should turn out to be insignificant is that the case we you see. The adjusted r square is still high because you still have included the right number of regressors, but the only point is you have over included the regressors here; now let us go and see; so now what has happened now?

Student: (Refer Time: 20:29).

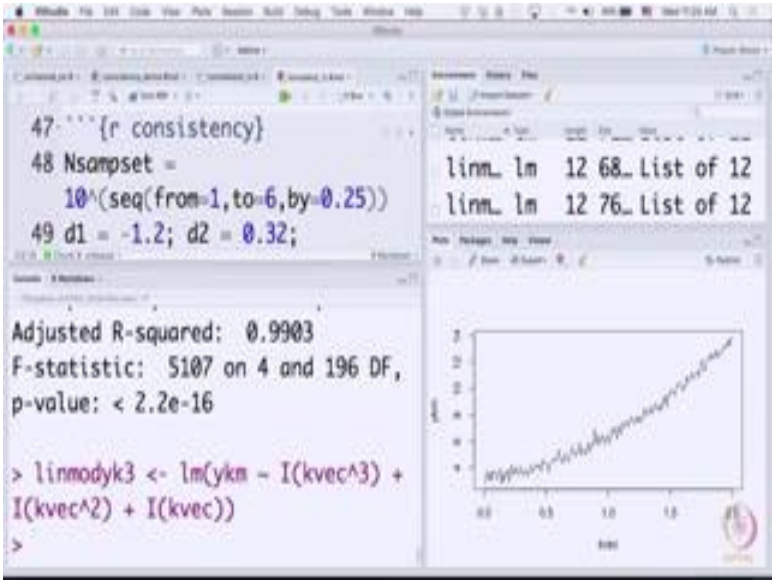
(Refer Slide Time: 20:35)



Everything has gone for a task, except the intercept term all the estimates have taken a beating right; you see that. So, what over fitting does is not necessarily only bring you large errors for the excess regressors, but can also spoil the game for the other important regressors as well that is why one should try and avoid over fitting. And the traditional approach is a bottoms up approach, what you mean by bottoms up approach is; you start with the a simple model when in doubt when you have no prior information; start off with a simple model based on your visual inspection, here the visual inspection suggests a quadratic that is why we started off with a quadratic; if that was also not clear we would have began with a linear one.

Sometimes it can happen that over fitting will show large errors for parameter estimates corresponding to unnecessary regressors, but there is no guarantee. What is definitely known is over fitting will bring in large errors in your parameter estimates, whether it is for all estimates or part of the estimates said that depends on the problem entirely.

(Refer Slide Time: 21:54)



So, for example, here if you were to go back and throw out the fourth order term right. So, suppose we do this and now ask for the parameter estimates still of course, adjusted r square remains very high, the model has done a good job of explaining it.

But now you see, it says that the intercept term of course, you know it is significant that is ok. It also says that the coefficients corresponding to k square and k 1 are significant, but the coefficient corresponding to k cube is not significant. In this case it has correctly told me that the excess regressor that I have included should be thrown out. Now the modeling is completed only when you go back and finish your throughout this additional regressor that you have included and develop your model with the right number of; the right kind of regressors.

In other words, you cannot actually just report the previous thing showed us, the previous thing shows us that you cannot say well I do not know what is the order of the polynomial; I am going to fit some 50th order of polynomial, it should automatically tell me which regressors should be thrown out. By telling me the corresponding insignificant parameter estimates; unfortunately that does not happen. It can happen that as you see here that is says all parameter estimates are insignificant, then you have to back track and back off and then say no maybe I over included so let me cut down, but that is the problem with a top down approach whereas with the bottom sub approach you know you are systematically adding.

Now, this is a traditional approach which has lasted for several decades; at least for about 60, 70, 80 years you can say. Now today there are methods which will automatically pick the regressors for you, under some conditions not all the times, but not so restrictive, even if you over parameterize there are methods which solve a different kind of optimization problem, they do not solve the least squares optimization problem, they solve what is known as a one norm optimization and there is a method called LASSO which stands for Least Absolute Shrinkage Selection Operator. It was developed in mid 90's by people at Stanford and that essentially selects the right variables for you. You may include some hundred variables in your regressor set and it will come out and tell you well only these regressors should have been included the other ones should not be included.

Of course provided some conditions are satisfied, we will not go into that. Now here of course, we are not talking of time series models, when we go to time series models things become a bit more complicated, but the basic concepts still remain. So, now you have seen that there are a number of calculations that are being made for example, sorry the standard errors are being calculated then the adjusted r square we have given the expression, but then there is a sum square; you know residual standard error that is being calculated and then I have also talked about distribution and so on.

So, let us turn to theory and ask how these calculations are made and then we will come back to an example on time series modeling, where I will demonstrate to you what kind of biases can result in time series modeling or otherwise; that means, you can have unbiased estimates as well, whether we get consistent estimates and so on in particular with auto regressive models. The reason for confining ourselves to auto regressive models when it comes to ordinary least squares is, we are discussing until now the linear least squares; what is linear here? The predictor or the y hat is linear in the unknowns and AR models if you recall they generate predictors that are linear in the parameters whereas, moving average models produce predictors that are non-linear in parameters. So, we are not in a position to apply least squares to estimating moving average models as of now because in a moving average model the predictor is a non-linear function of theta, only when we discuss non-linear least squares we will be in a better in a position to talk about moving average models.