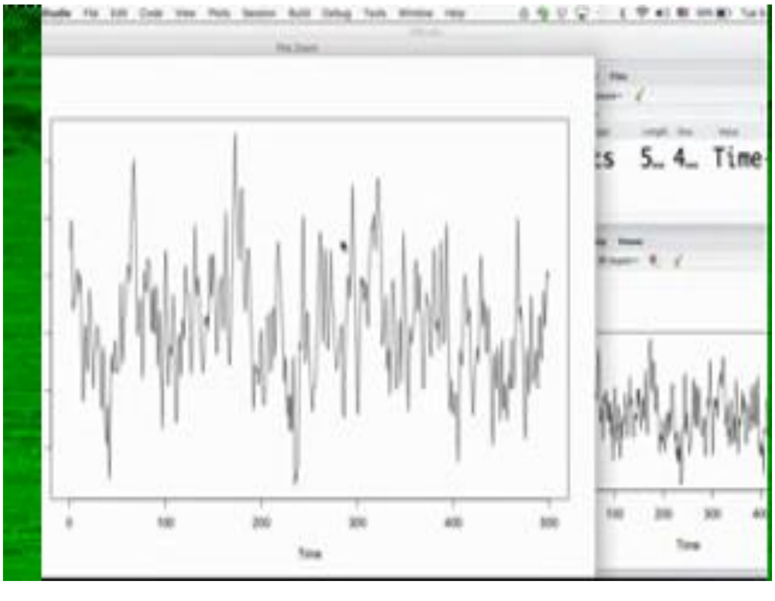
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 50
Lecture 22A - Models for Linear Stationary Processes with R Demonstrations 14

So, let us begin with simple demonstration of how ARIMA models are fit in our, I will show you later on how to simulate, but let us say I am given a series v k, I do not want to show the simulation part right now because I just want to make it as realistic as possible, so that the real model is hidden from you. I have a now the r studio with me and fairly simple set of comments you probably aware of most of them by now. So, what will do is we look at the series we can say always say when you are given a time series do not and you are ask to build a model, just do not straight jump to ACF, PACF, ARIMA model and so on. The first thing that you want to do is you want to really break this with the data and you want to therefore plot the series.

(Refer Slide Time: 01:06)

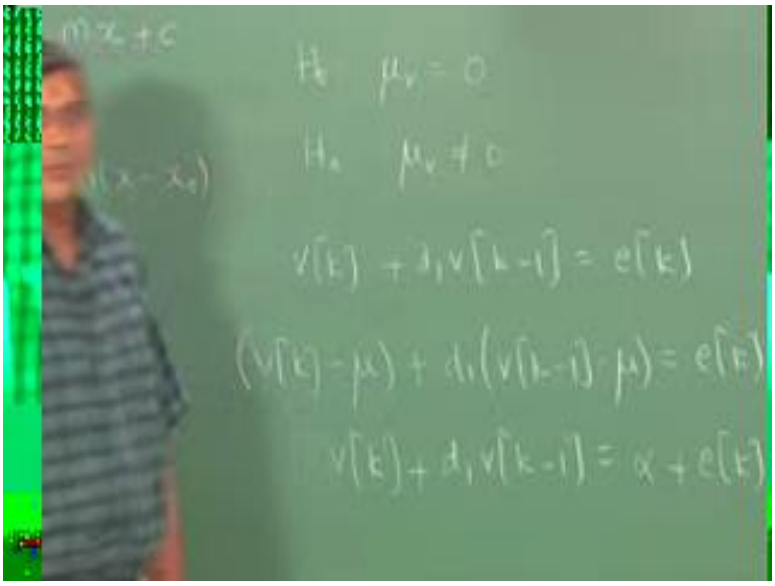


So, let us do that as a first thing and here you have the plot I am going it is zoom it out so that you can see better. The reason for visually examining the series is to see if there is anything that you can just spot by the eye for example, some trends or oscillatory features or out layers and so on and that is actually really useful because certain decisions that one has to make and also certain validations that then has to make in the course of time series modeling, really hinge on your ability to visually observe to do that and your ability to visually observe something comes about only if you plot the series. So, it is a very very good habit even if you have simulated the series. In this case I have simulated the series and even then I am looking at the data, because a lot of times we may have some model in our minds and we may end up instructing something else in r so that this series is completely different. So, you want to really make this habit in not only in this course, but in all exercises of data analysis that you plot the series.

So, when you look at the series here, nothing really straight away tells us that there is a trend or that there is an oscillatory feature and so on and these only beginning right; if you have any kind of inkling that there is non stationarity in this for example, an integrating effect and so on, then there are unit root test available which we shall talk about later on; at this moment we will say you know as a preliminary thing I do not see anything non stationary here no trend and so on, I will assume stationarity and proceed. Generally if you make a wrong assumption and if you are careful and doing a systematic analysis, at some point you will figure out that you made a mistake, So not to worry so much.

So, let us actually get back and now generally the next step is to do some kind of a simple non parametric analysis look at the mean, variance and so on. So, will ask for example for the mean of it very simple statistics, not that I am going say yeah you know the 0.23 is very familiar and so on nothing like that; is just to get a feel of the data of course, when you look at this is sample mean that we have computed it is not the theoretical one, now at this stage one can set up a hypotheses test asking for the test where you test for 0 mean, you can set up a 0 mean hypotheses test.

(Refer Slide Time: 04:00)


mx + c
H0 μv = 0
Ha μv ≠ 0
(x - x0)
v[k] + d1 v[k-1] = e[k]
(v[k] - μ) + d1(v[k-1] - μ) = e[k]
v[k] + d1 v[k-1] = α + e[k]

In other words you can conduct hypotheses test with the null hypotheses as mu v is 0 against the alternative that it is in and of course, we have not gone through the formal hypotheses test, but the videos that you the video lectures that you been hopefully listening to do tell you how to conduct hypotheses test, we will go through that when we discuss estimation theory.

So, 0.23 say I may seem quite high on the face of it and that is where the hypotheses test is going to come to your rescue and help you conduct a systematic test rather than just looking at the face taking things on the face value. If we are a layman as a layman we will say 0.23 is fairly (Refer Time: 04:49) may be the mean of the series is not zero, but at this moment we will not go through hypotheses test, we will through the help of modeling even be able to answer this question.

In general I also want to tell you until now we have been looking at models for processes with 0 mean for example, if I have written AR 1 model, then I have written this kind of a difference equation and this is fine as long as the process is of 0 mean; suppose the process is non 0 mean then the and of course stationary, we have still remain in the stationary then the model changes to this.

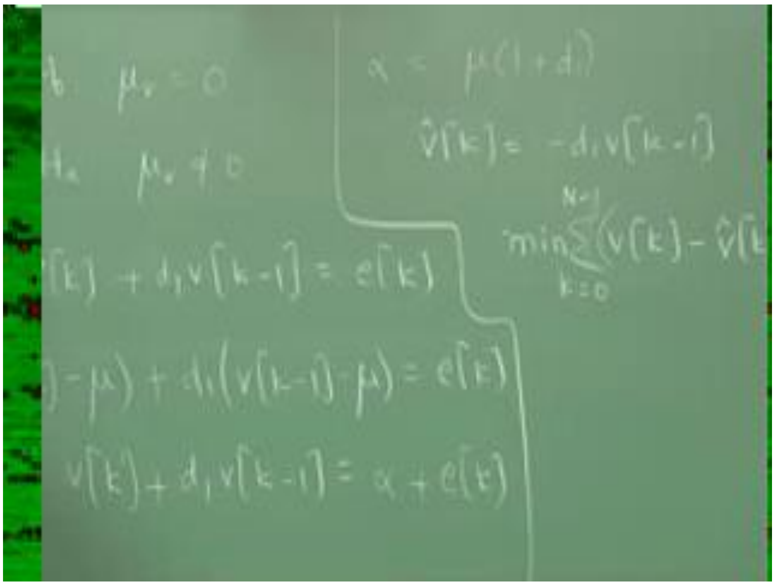
In other words, instead of building models for v k, you would be building models for v k minus mu departures from the mean; and that is a standard thing even in deterministic world for example, in the deterministic world if I have let us say some relation between y and x as y equals m x plus c; is this a linear model? Initially lot of them lot of I heard lot of yes whispers, it if you go by the traditional je coaching type thing look at the power of x and so on then it appears linear, but from in a strict sense it does not because it does not satisfy the basic requirements of a linear relationship a linear mapping, which is the principle super position or even homogeneity. So, x if it gives you y, alpha x does not give you alpha y; so, in this case we do not call this as a linear we call this as an a fine relationship, which can be transformed into a linear relationship by constructing what is known as a model in terms of deviation variables.

(Refer Slide Time: 07:01)

So, if I have a reference point x naught y naught in the two dimensional space of x and y then I could rewrite this model as y minus y naught equals m times x minus x naught ok.

So, in such a case as you can see what we are building is linear model, but in terms of deviations of y from y naught. So, you can also think of this model in this way; e k still remains as a 0 mean right, the other way of looking at this model is to rewrite this model in terms of this, but add an intercept term. So, you could say that there is some alpha plus e k. So, you are now transferring the non 0 mean nature of v to as if you have now and in case, you can think of this alpha plus e k as new e k with non 0 mean equaling alpha, but this alpha has a relationship with d 1 and mu and what would be the relationship? So, what would be the alpha here for us? D 1 times or mu times one plus d 1;

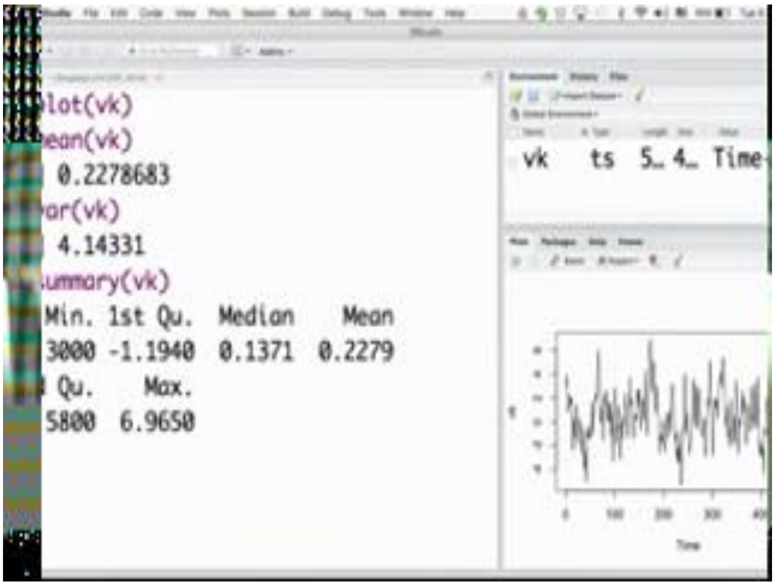
(Refer Slide Time: 08:31)


H_0: $\mu_v = 0$
H_a: $\mu_v \neq 0$
$[k] + d_1 v[k-1] = e[k]$
$) - \mu) + d_1(v[k-1] - \mu) = e[k]$
$v[k] + d_1 v[k-1] = \alpha + e[k]$
$\alpha = \mu(1 + d_1)$
$\hat{v}[k] = -d_1 v[k-1]$
$\min \sum_{k=0}^{N-1} (v[k] - \hat{v}[k$

So, it does not matter essentially what this tells us is if you have a process with non 0 mean, either you think of this model; that means, you subtract the mean and then fit a model or you fit a model with an interceptor it usual model with an interceptor and this is important because when I pull up the help on this routine which estimates the ARIMA model for you, it will ask a few questions from you; do you believe the mean is 0 or not and so on. So, you have to supply some inputs. So, using those options we will be able to answer the question for this series whether this null hypothesis holds or not.

So, now we can also look at variance of v k not there it is going to tell us a lot just you know for information sake. In fact, we can ask for some summary statistics right which gives you the minimum value, the maximum value, the mean median and so on and one of the things that I have also said let me see if you can give the correct answer with regards to this question of what we should check for before even proceeding to building an ARIMA model or it is a linear model that we have fit in?

(Refer Slide Time: 09:59)



```
lot(vk)
ean(vk)
 0.2278683
ar(vk)
 4.14331
ummary(vk)
Min. 1st Qu.  Median    Mean
3000 -1.1940  0.1371  0.2279
 Qu.    Max.
5800  6.9650
```

Is there anything more that I should do before jumping to computing sample ACF and sample PACF?

Student: (Refer Time: 10:15).

Sorry.

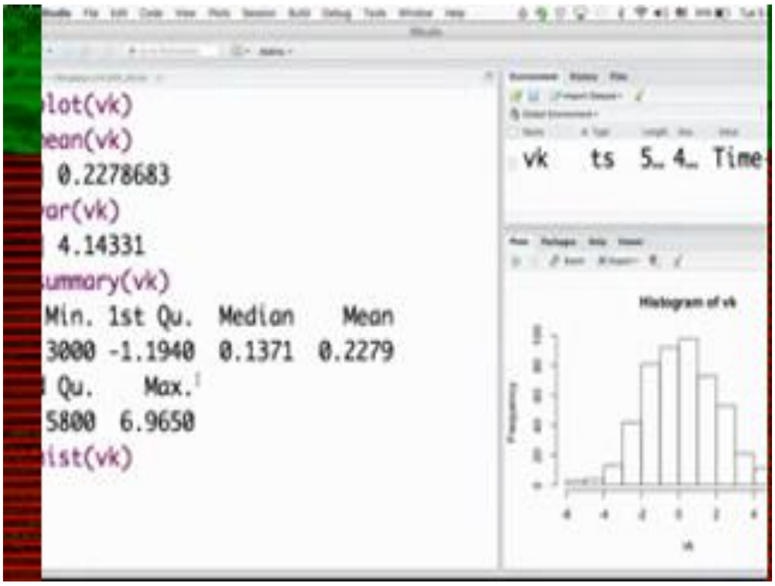
Student: (Refer Time: 10:16).

I am sorry any answers is that anything else that I need to look at, any assumptions that we make on building linear models? Stationarity is anyway something that we have said visually I do not see any evidence at the moment.

Student: we want (Refer Time: 10:47).

Correct. So, that is something that we assume Gaussianity is not necessary for building a linear model, but necessary for the optimality, in the sense if the model that you are going to fit is expected to give you an optimal prediction, optimal in the sense as good as your conditional expectation then you are assuming joint Gaussianity. So, you want to be able to see if the series has some kind of a Gaussian distribution or not; obviously, we are working with the single realization you cannot really do much, but there is something that you can do without instead of just skipping that steps altogether. So, we can look at a histogram at the histogram of the series and ask if the distribution looks Gaussian.

(Refer Slide Time: 11:31)



```
lot(vk)
ean(vk)
 0.2278683
ar(vk)
 4.14331
ummary(vk)
 Min. 1st Qu.  Median    Mean
3000 -1.1940  0.1371  0.2279
 Qu.    Max.
5800  6.9650
ist(vk)
```

vk ts 5.. 4.. Time

It is a sample distribution that we are constricting, it is from a single realization and I think all of you are able to see the plot right or I can actually zoom it out for you so that you can.

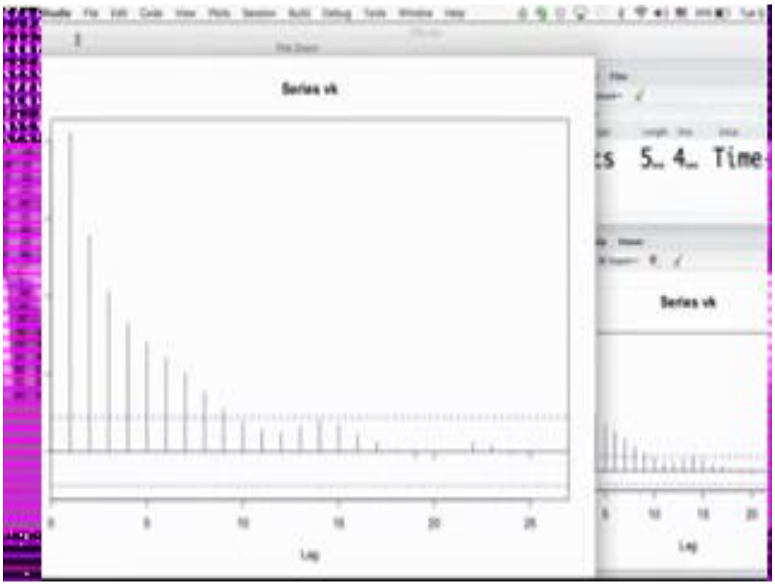
So, here is the sample histogram; that we generated from the given realization and there is a strong indication that it is it has the these observations have a Gaussian distribution of course, again here one can go through a formal test of Gaussianity there are Gaussianity test we will not go through that, but you want to make note of these tests sorry these steps and again write against each step that some steps we are just doing in an informal way and later on we will do the most important steps very rigorously.

So, since we are more or less convinced that it is stationary by visual inspection, Gaussianity at least through histogram calculations, it is fine even otherwise you are to fit an ARIMA model please keep that in mind, just because you find that non Gaussian distribution you should not completely apply the brakes and close your laptop and enjoy open your phones. So, do not do that; you can still fit an ARIMA model get decent forecast and so on.

So, let us now get into the time series aspect of it. So, now, you want to see if there is any scope for predictability where we turn to ACF right and here I am using the ACF routine in the stats ah package and of course, here we can ask for a maximum lag of something like 20, 50 and so on 30, 50 and so on will just use the defaults, by default it plots the ACF and there is an option that you see I do not know how well you can see here it says among the many options there is something called d mean equals true.

Unfortunately, I do not know how well it is visible to the people sitting in the back, but there is an option and that option has got to do with whether you want to use the standard defamation of autocovariance function where we subtract the mean or the signal processing the engineers definition of the ACF as we do not subtract them here and sometimes it can make a difference, but here we will stick to the standard one as I said throughout the course will use the standard one and the default is little plot correlations for us.

(Refer Slide Time: 14:17)



So, here is the ACF I hope all of you are able to see this plot and this is where if we had made a mistake on stationarity assumption for example, remember non stationarities of are different kinds not all non stationarities can be pick by the ACF, there are different types of non stationarities: non stationarities in the mean, non stationarities in variance; within non stationarities is in mean you have trend type which are deterministic type non stationarities and you have the stochastic ones which lead us to the random work type or non stationarities and so on, but among the variance non stationarity is the variance can change with time for the process and your ACF may not be able to pick that up; it may still show you at decaying kind of ACF.

So, relaying on ACF for checking for non stationarities is a good thing to do, but one should keep in mind that it does not detect all types of non stationarities. So, if there is a trend type or if there was an integrating effect one could easily expect to slow decay of the act, but that does not seem to be the case here and at this stage if I look at the ACF what are the inferences that I can draw? Any obvious ref inferences that just you know just speaking to you, ACF is telling you look this is what the series can contain or can be thought of why is so different?

Student: (Refer Time: 16:06).

So, the question that comes to mind is because you are doing this for modeling, first of all this is series have predict predictability, it clearly has right the estimates are way all the at least there is one estimate which is significant at non 0 lags, that is the first point. Second point is now we want to ask which model is suited; whether it is MA or AR and that is where we start looking for certain signatures and what kind of a signature do we see here?

Student: (Refer Time: 16:37).

AR model exponential d k; anybody wants to debate that? No it could be moving average model of high order no possible right because after all these are estimates and they do not carry the exact theoretical signature so that is the ambiguity with which we are always resented when it comes to time series modeling and then one has to carefully explode different options and come up with the suitable model. So, at this stage if I have to think of a moving average model for this process, then what would be an appropriate order to start with?

Student: 979 (Refer Time: 17:19).

9. So, we have navagraha. So, we have here 9th order excellent, but if I have to now think of this as an AR because you know maybe I can think of the decays exponential then I can turn to psc, because 9th order seems a bit high, generally speaking you should be able to model a lot of stationary time series processes with fairly low order ARIMA or AR and MA and so on unless you know have some special series that are waiting for you during may be exams and so on but.

So, let us turn to PACF and ask what story it has to tell us. So, this is the PACF for you and the PACF tells us some other story and it says yeah you can think of this as an AR model of.

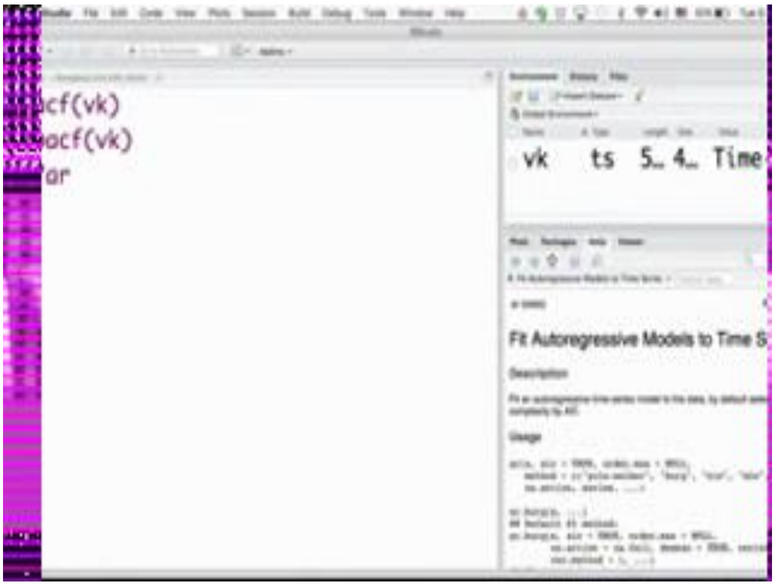
Student: (Refer Time: 18:14).

Fifth order right. So, in terms of parsimony, AR seems to be the winner you know with respect to ma model, but we have another choice two which is ARIMA and as I have said none of this ACF and PACF truly tell me what could be suitable order if I were to fit an ARIMA model. ARIMA model orders are generally found by trial and error in a systematic way. So, at this stage instead of fitting an MA 9th and checking out we can actually start of with AR 5 because there is no other information available to be and when there is no other information available generally the preference is for AR models because of ease of estimation and so on and in this case mainly the principle of parsimony is telling us that a fifth order AR is perhaps better suited.

In some other case it may be that the AR model is 10th order AR model actually is suited to the series visa v and MA 2, then you my may want to pick up MA 2. So, that is option number 1; we can fit an AR 5, but given that it is AR 5 and MA 9 and so on there is a chance that the underline series is actually coming out of an ARIMA process correct and because we do not have any indication of what is the suitable order, we start off with ARIMA 1 1. So, there are two models that we are going to fit: one is an AR 5 another is an ARIMA 1 1.

So, let us do that now when it comes to fitting AR models as I told you already because AR models give raise to linear predictors and therefore, it is much easier to estimate the parameters, there are special routines and we have talked about Yule walkers method, we have talk we have spoken of Levinson Durbin’s algorithm and so on and then there is burgs method and so on.

(Refer Slide Time: 20:22)



```
cf(vk)
ocf(vk)
ar
```

vk ts 5… 4… Time

Fit Autoregressive Models to Time S

Description

Usage

There is a separate routine for estimating AR models and the routine is simply AR, you can pull up the help on AR, I m sure you are at least in this room you are unable to read the help unless you have a magnifying lens in front of you, but unfortunately I do not know how to increase the font size in the help, but if you look at the AR command of course, apart from asking for the series, it is asking for an important piece of information from the user what is the order that you want to fit.

And let me read this for you here, it says that apart from supplying the series, the second option is AIC. AIC stands for Akaike information criterion it is an information criterion; information theoretic measure derived by Akaike, using information theory concepts and the it is a Boolean variable setting AIC equals true allows the ma routine to pick among a range of orders, a pick model with a suitable order among a range of orders that the users specifies we said fifth order is good, but it is only a good guess we do not know that is the final answer.

So, may be as a user I would think 5 to 7 is what I want to explore and I can specify that through the third argument, which is listed as order dot max and again if you set AIC true then it use a explores all orders from one to the maximum specified order; If AIC is set to falls it will freeze the order to the maximum order that you have specify and then there are other options here for example, it is asking for the method that you want to use whether you want to use Yule walker method or you want to use burgs method or ordinary least squares and so on.

Right now let us not worry about the methods when we move to estimation theory will know what are all this method, we will stick to the default one may be default one I think is Yule walker. And then na dot action. So, many at times your series can have missing data and na stands for not available. So, I asking if you give me a series with missing data what do I do with that, do I actually get scared and start working or do I just ignore it or do I just linearly interpolate omitted that one has to specify right now that is some not something that we should be worried about and then there are other options that you want to see that you can give and if you want actually get a full list it is given here there is an option for d mean and that is important; that means, should I estimate the mean or not if I set d mean equals true d mean means removing the means. So, if you whether you want to fit this model or this model. So, if I said d mean equals true it is going to fit this kind of a model and get you an estimate of I believe the alpha. In fact, it is unfortunately and this is one of the things that was there is also discussed on stofa us website on some of the fallacies that you have with this routine and the terminology that is used in r.

So, although it says it going to fit a model like this ultimately it is going to report alpha, but does not matter in the sense what you want to know is whether first of all alpha should be in the model or not because if alpha is 0 mu is 0, because is the relation that we have between alpha and mu, that is the first thing of concerned to me; if it is not then from alpha we can always figure out what is mu because the model is the routine estimates d 1 for me right. So, we will we set d mean equals true and then let the algorithm tell me what it does.

Now, we can do that. So, let us actually the then there are other options do not worry about the remaining options and so on and in every when you look at the help on any routine, also understand apart from what it needs, once it is done what it returns and that is usually described under value. So, will it is telling you what are all the things that is actually going to come back with and it is going to come back to you with of course, the AR coefficients, the order that it has selected of course, if you set if you set AIC equal false, order is going to be the same as you have supplied as order dot max and then it is actually going to give you the variance of the prediction error, it is going to report says x dot mean it is going to report the mean that it has a estimate. In fact, it is going to report alpha and then it is going to tell you how many observations were used in the time series that may be a bit of a surprise to us and supplying let us say in this case I have 500 observations right that is something that should know I would not spoken about it, but you can always look up the data sorry and I should tell you. So, there are finite observations that I am feeding in you should use of finite observations, what you mean by number of observations used in the series unfortunately there is not the case.

Remember your order will tell you how many observations have to be thrown out depending on the method that you are using and it may seem strange why I have to throw out the few observations. So, if you look at the simple example here let us say it is an AR 1 that you going to fit in to the series, when I set up let us say the linear regression equation I am going to set up the regression equation at every observation instant that I have and then straight forward, standard linear least squares solution and so on. So, you will be when we say setting up the linear regression equation, what you are going to do is you are going to write the predictor form and require that the predictor or the prediction be driven as close as possible.

In this case is going to be a least square if least square approaches used, this is what you are going to do. So, you have v k minus v hat of k the whole square the sum of that being minimized; this is not the only way, but you should understand the underlying principle the underline principle is I have a predictor that I have written using you know assuming it to be AR 1 and d 1 is the decision variable and I want the algorithm to drive d 1 to tune d 1 such that v hat of k is taken as close as possible to v k; now as close as possible is a qualitative statement, in the least squares approach what we are saying is drive it in the sense of least Euclidian distance right and since we have n observations, we want actually to make sure that all observations predictions are as close as possible in the least square in this Euclidian sense to the respective observations.

And since we cannot afford to just please one observation, you want actually satisfy all observations; this kind of an approach is taken that is a idea behind least squares. Now I have written here the summation running from k equal 0 to n minus 1, assuming of course, that our time index always starts from 0 although in r it starts from 1. Now you should straight away notice that one cannot construct this summation strictly speaking from k equals 0 why.

Student: (Refer Time: 28:37).

Yeah. So, we do not have to; I have v 0, but I do not have v hat of 0 and there is a whole lot of research done on how to handle that. So, about I think about 10 PhD is and few masters and so on, how to handle this; lack of initial conditions. So, the best thing is it is a b tech approach is throw away, forget it why I do not want to do any PhD, just give me this s grade and in this course I will get my best job in (Refer Time: 29:10).

So, k equals 1 to n minus 1 I have thrown away, I said why unnecessarily create a problem and then try to solve spend beards and mustards and so on grow them and so on I do not need any of those. I will start my prediction from one onwards, now I have everything available and that is what exactly this AR is routine is asking; how many observations were used of course, you know that depends now on the order if I have to use a 5th order, I will have to throw away five observations and that is one of the things drawbacks with this kind of an approach as you fit higher and higher orders, you will be throwing away that many observations and that is not good from an estimation view point particularly when the number of observations is not large, each observation is valuable it has some information throwing away is not such a great idea.

But when you have large observations always you know when anything is in excess we become very intelligent and we say does not matter [FL] no problem. So, that is the approach we are taking and the number of observations is exactly that it says how many observations it used in the series, but again let me tell you this is only one of the many approaches that are available and then it reports for example, the method itself that used and the residuals that is very important right, when I fit a model it I want to ultimately know whether I have done a good job of modeling and residuals hold the key. So, this routine not only estimates the parameters for you, but also the residuals good.