Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 05

Lecture 21: Forecasting Basics

Hello all, welcome to this new lecture in the course on time series modeling and forecasting using R. So again, we're starting with a new week, and the idea throughout this week will be more on forecasting. So the title that you see here comes after you have a time series model, and then you fitted all the unknown parameters using the estimated values and so on and so forth. And then, really, the idea of forecasting is how do you take the model into the future? Or, in other words, how do you predict something in the future? Okay, so depending on a historical set of observations, how do you propose the set of values for a futuristic-looking series, okay?

Now again, often people have small discrepancies when it comes to all these alternative terms. For example, estimation, then prediction, and lastly forecasting. So I am very sure that many of you might have sat through a regression course. So in a regression course, we focus more on the middle aspect here, which is prediction. But if you are sitting in a time series course, then the alternative terminology for prediction would be forecasting.

But again, nevertheless, you have subtle differences between all three aspects when it comes to, let's say, gathering some historical data and then proposing something in the future. By the way, one of the most important objectives in a time series analysis, or when you are analyzing any practical data, is to forecast its future values, alright. So, generally, the process is that you gather some historical data and then you try to forecast the same time series for a certain length in the future, okay. Now, again, what exactly is the difference between these three alternative terminologies? So, estimation.

So, estimation is nothing but estimating the unknown parameters in a model. So, let us say you have created a statistical model or, let us say, a very basic time series model. For example, let us say AR1 or MA1, etc., right? So, regardless of whatever model you have created, the idea behind estimation is nothing but estimating the unknown parameters in the underlying model, right? Just to give you a very quick example, let us say AR1, how many parameters are involved here?

So, probably you can pause the video and then just think in your head and then kind of come to an answer as to the AR1 model contains a single unknown parameter, which is, let us say, phi 1 for that matter. So, estimating the unknown parameter, which is phi 1, is nothing but called estimation. Now, what about prediction? So, prediction means the value of the random process when we use the estimated values of the parameters. So, let us say initially you have a set of all the unknown parameters, then using some estimation technique, you try to estimate the unknown parameters.

So, let us say maximum likelihood or method of moments or maybe least squares, and so on and so forth, and once you have estimated all the unknown parameters and when you replace all the estimators back instead of the actual unknown parameters, that idea or that exercise is called prediction. Now, again going back to my same example of the AR1 model structure. So, let us say if you write down the AR1 model structure, it is something like this. So, let us say y t equals some constant plus phi 1 y t minus 1 plus probably some error, alright. So, we start with a very basic AR1 structure where the current y t sort of depends on a single historical data point, which is y t minus 1, plus some random error, alright.

$$Y_t = \phi Y_{t-1} + a_t$$

Now again, as said in the beginning, the estimation aspect involves estimating this unknown parameter phi 1. But once you estimate the unknown parameter phi 1 and the moment you replace it back in this model structure, something like phi 1 hat. So phi 1 hat is nothing but the estimator of the unknown parameter phi 1. Now this exercise is called prediction. Because now nothing is unknown that you see on the right-hand side.

$$\hat{\phi} \text{ (Estimate of } \phi\text{)}$$

Because phi 1 hat is now known, which comes from the data. ET could be generated using some random process. So let us say normal distribution or maybe some other distribution, and so on and so forth. Because ET is nothing but white noise or ET is random error. So in a sense, nothing would be unknown once you replace the estimator back in terms of the unknown parameters.

$$\hat{Y}_t = \hat{\phi} Y_{t-1} \text{ (Prediction)}$$

And then this intermediate exercise is called prediction. And now lastly, what exactly is forecasting? So the value of any future random process which is not observed by the sample. So, let us say you work with a fixed set of observations, which is called a sample. Then you try to estimate the unknown parameters; that exercise is called estimation.

$$\hat{Y}_{t+1} \text{ (Forecasting)}$$

When you replace back the estimators in terms of the unknown parameters, that exercise is called prediction. And using this estimated model or using this predicted model to predict something in the future. That exercise is called forecasting. So hopefully, this slide is kind of important as it tells you more about the differences between the three aspects of analyzing time series data, which are estimation, prediction, and forecasting. Now again, throughout this week, we will delve deep into some ideas about forecasting.

So let us say smoothing techniques. So simple smoothing, exponential smoothing, Holt's approach, Holt and Winter's approach, etc. And obviously, towards the end, we will try to bring in a practical data set and then try to connect all the theory that we study in this week. Now, just to give you a brief background about what path one has to follow when it comes to, let us say, estimation, then prediction, and then ultimately forecasting, okay. Now, I think this is the exact same example we took up earlier.

So, let us say this is the famous AR 1 structure, right? And then again, phi here is unknown. So, what do we do? So, we try to estimate that using, let us say, maximum likelihood or method of moments, etc. Now, once you estimate this unknown phi, we can write down phi hat. So, phi hat is nothing but the estimate of the unknown parameter phi.

And the moment you replace it back in the same equation, so instead of phi, if you write down phi hat, then on the left-hand side, I can write down y hat, right? Because this y hat is nothing but my predicted y. So once you replace the unknown parameter using its estimators, then the left-hand side, which is essentially nothing but the yt value, could be predicted. And lastly, if you want to extend that into the future, so let us say yt plus 1 hat, this exercise is called forecasting. So estimation, prediction, and then forecasting, they kind of follow a certain sequence in that sense.

Okay. Now, initially, we will talk about one particular structure when it comes to forecasting. So, let us say we will start with forecasting from an ARMA model initially, and then this technique is called the minimum mean squared error forecasts, right? Now, again, if you are not very sure as to what you mean by mean squared error, so probably you have to pause the video and then probably revise exactly what you mean by mean squared error or, in short, ARMA.

What we call MSE, right. Now, again, the idea behind what exactly is MSE is widely studied in some regression courses when you try to minimize the MSE while fitting, let us say, a linear regression or non-linear regression, etc., okay. But in other words, MSE is nothing but the deviation of the actual values and the fitted values or the predicted values, and then it has a square term on that. And hence the name minimum mean squared error. Now, what exactly do you mean by minimum mean squared error forecasts or minimum MSE forecasts?

So, again, we will take up a small example. So, let us say, using an observed time series y1, y2 up to yn. Now, again, here I might change the notations very slightly. So, instead of writing down these values in terms of capital letters, I will use small letters because small letters kind of denote a generated sample, right? Or more like an observed time series. So, observed time series is still a random variable, but then just to sort of differentiate between this capital notation and the small notation.

So, small notation means that once you observe a time series. So, let us say you collect some practical data on temperatures or rainfall, etcetera, and then, of course, the sample size has to be n, but then we can replace the capital letters by small letters, okay. Alright, so using an observed time series y1, y2 up to yn, the question is, how do we forecast some future value? So, let us say yn plus 1 or yn plus 2 or yn plus 3, etc. And, by the way, here this sample size, which is nothing but n, is also called a forecast origin, right?

Because you already have this data y1, y2 up to yn, and then the exercise is to forecast something in the future. So, you keep or you kind of fix n as the forecast origin, and then n plus 1, n plus 2, n plus 3, and all the subsequent subscripts would be nothing but the forecasts. And then here I have jotted down a few examples of the initial one. So, let us say y n and then within brackets 1 with a hat on top is nothing but the forecast value of y n plus 1. Now, again, imagine this in your heads that you are observing this series.

So, y1 y2 up to yn then I will sort of create a partition here and then all the subsequent values. So, yn plus 1 yn plus 2 etcetera have to be forecasted right. So, a forecast of yn plus 1 would be nothing, but yn plus 1 hat a forecast of yn plus 2 would be nothing, but yn plus 2 hat. Or in other words, these are some alternative notations. So, one can actually write down y n plus 1 hat as nothing but y n hat and then within brackets 1.

Or I can actually write down y n plus 2 hat as nothing but y n hat and then within brackets 2. etcetera ok. So, yn hat within brackets 1 is forecast value of yn plus 1, then this guy is the forecast value of yn plus 2 or in general this guy where you have l in the brackets is nothing but the forecast value of yn plus l. So, the last thing is a more generalistic kind of a version. So, this is called as an l step ahead of forecast. And by the way, the first one is called as one step ahead forecast.

The second one is called as two step ahead forecast. Then you will have a three step ahead forecast, etc. Or in general, we have this common notation, which is being followed by a lot of practitioners. And how exactly are you going to find this L step ahead forecast is sort of we will exploit this technique of minimum MSE criteria. Alright.

So hopefully, the idea of how you denote a forecast and then what exactly is the difference between the notation of the actual observed time series and the forecasts should be clear. Okay, so again, forecasting from an ARMA model. Now again, unfortunately, the first set of slides this week will be slightly technical because when it comes to forecasting, you have to understand the nitty-gritties of all the involved equations. But I will try to keep all the things simple and try to explain each and every point in a bit more detail wherever possible. So, again, forecasting from an ARMA model.

Now, again, let us say if you want to create an L-step ahead forecast. Now, again, notation-wise, this is yn hat and then within brackets L. So, this L-step ahead forecast can be treated as nothing but the conditional expectation of what? So, the conditional expectation of this random variable, which is yn plus L, given all the data points that you have, which is nothing but yn, yn minus 1, yn minus 2 up to y1. All right, isn't it? So the data points that you have or the observations that you can actually use are nothing but the observed time series, which is nothing but Y1, Y2, Y3 up to Yn only. Right.

$$\hat{Y}_n(l) = E(Y_{n+l}|Y_n, Y_{n-1}, \dots, Y_1)$$

So, given this information or given all the information contained in the first n observations. We have to somehow come up with an idea about the L-step ahead forecast or the L-step ahead value, which is denoted by Yn plus L. So, can we not write it down as some conditional expectation of that particular value given whatever information we have? So, essentially, the L-step ahead forecast is nothing but the conditional expectation of Yn plus L given the observed sample or given the observed data. So, if you are able to kind of digest this idea that the L-step ahead forecast or 1-step ahead forecast or 2-step ahead forecast is nothing but some form of an appropriate conditional expectation. Now, how do you evaluate that conditional expectation? That entirely depends on what model you are kind of assuming.

So, again, forecasting from an ARMA model. So, revisiting the ARMA model for a second. So, this is the famous ARMA model. So, let us say y_t equals some constant theta naught plus this is nothing but the autoregressive structure up to this point, and then this entire thing is nothing but the moving average structure. And hence the model is ARMA.

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

So, autoregressive and then moving average, and of course the orders here are P comma Q. So, this is a very general kind of an ARMA model having orders P Q. And by the way, all these are the unknown coefficients. So, phi 1, phi 2 up to phi P, and similarly in this MA structure, you have theta 1, theta 2 up to theta Q. Now, probably we have discussed this in one of the earlier lectures also, that I can actually write down this entire ARMA structure in a slightly concise form by utilizing some coefficients, ok. So, what I will do is I will combine this entire AR structure in terms of phi p B coefficient applied on y_t equals theta naught plus now I will combine this entire MA structure containing all the theta coefficients into this theta q B coefficient applied on a_t, ok. In other words, think of a situation where you bring this entire AR structure on the left-hand side, right?

$$\phi_p(B) Y_t = \theta_0 + \theta_q(B) a_t$$

And then keep this MA structure on the right-hand side. So, essentially what is happening is I can actually write down an AR coefficient written down as something like phi p B applied on y_t equals the constant which stays there, which is theta naught plus some concise coefficient in terms of thetas, which is theta q B applied on it. Now, again, what exactly are these coefficients, what exactly is capital B? So, hopefully, all the answers should be there in some of the earlier lectures. For example, B is nothing but the coefficient which, if you apply B on y_t, it is nothing but y_t-1.

So, it is nothing but the backshift operator. So, I can actually write down all these subscripts or all these intermediate terms involved in both the AR term and the MA term in terms of some backshift operators. Isn't it? Because yt minus 1 is what? So, yt minus 1 is nothing but b operated on yt. Okay.

Further, what is y t minus 2? So, y t minus 2 is nothing but b operated on y t twice. Alright. So, or rather b squared applied on y t and so on. Okay.

So, each of these terms, if you see, are nothing but some powers of the backshift operators applied on y t. Can you see that? And similarly here. Okay. So, in other words, I can actually write down both these structures in concise kind of polynomials, which are nothing but coefficients. Okay.

Alright, and then the idea when writing this model, the ARMA model in this structure is it will make all the forecast writing slightly easier, right. Now, the actual L step ahead value of the process is given by this guy, right. So, what I will do is simply replace T with N plus L in the ARMA equation. So, can you do that? So, if you replace T with N plus L, this is exactly what you will get right.

$$Y_{n+l} = \theta_0 + \phi_1 Y_{n+l-1} + \cdots + \phi_p Y_{n+l-p} + a_{n+l} - \theta_1 a_{n+l-1} - \cdots - \theta_q a_{n+l-q}$$

So, theta naught is still there then phi 1 and then y n plus l minus 1 up to phi p y n plus l minus p and then subsequently all the errors. So, instead of a t you will have a n plus l. Now, instead of minus theta 1 a t minus 1 you will have minus theta 1 a n plus l minus 1 etcetera. So, to sort of get the actual L step ahead value of the process, I will simply replace T with N plus L. So, now looking back at the form which was written down in a concise form. So, if you go a slide back. So, using this structure, what would happen if you bring this phi p b on the right hand side, right?

So, what would happen again, if you bring this phi p coefficient in the right hand side, then you will have a ratio of the coefficients, right? So, there will be theta b divided by phi b, is not it? If you bring this phi b on the right hand side, and then this is exactly what is written down on the next slide. So, considering the random shock form, so the following equation that you see is also called as random shock form of the same ARMA series or of the same ARMA model. And by the way, what we are doing is we are bringing the phi coefficient on the right hand side, and then it has to go in the denominator, right.

$$Y_{n+l} = \theta_0 + \frac{\theta_q(B)}{\phi_p(B)} a_t = \theta_0 + \psi(B) a_t$$

$$= \theta_0 + a_{n+l} + \psi_1 a_{n+l-1} + \psi_2 a_{n+l-2} + \cdots + \psi_l a_n + \psi_{l+1} a_{n-1} + \cdots$$

So, we can write down something like this. So, y n plus l equals theta naught is a constant plus theta b coefficient or sets of coefficients divided by phi b coefficients applied on a t. Or I can actually assume one common coefficient or one common notation for this ratio, let us say psi. So, essentially y n plus l is nothing but the constant theta naught plus some other coefficient or series of coefficients which is denoted by psi b operated on a t. Okay.

And now I can expand this, isn't it? So, psi b is nothing but I can get down all the coefficients which are involved in psi b in terms of let us say psi 1, psi 2, psi 3, up to let us say psi l, then psi l plus 1, etc. Okay. Just note one thing here that you will not have any endpoint to this expression. So, it will extend up to infinity.

Okay. Hence, I can write down theta naught plus a n plus l plus psi 1 a n plus l minus 1 plus psi 2 a n plus l minus 2 dot dot dot and then psi l a n then psi l plus 1 a n minus 1 plus etc. So, essentially what I am doing is again if you go back just for a second, the first step is I have brought this phi coefficient in the right-hand side, it goes in the denominator and then we have come down to this random shock form of the series and then I am doing nothing but I am giving this ratio a slightly different notation which is psi b and then expanding. Now, here since psi b is applied on a t, you have to expand all the subscripts in terms of the errors. Now, one more point to note here is that if you take a look at this notation AT, so use a slightly different notation here, because later on we will require the ET notation, which we have been following for denoting errors to denote something else in this forecasting literature.

But again, AT is what? So, AT is nothing but the error vector in this case. So, I am using a slightly different notation than ET to denote the errors, which is AT. Alright, now what? So, once you write down the psi b coefficient applied on AT, then the last thing we would do is we would take the conditional expectation to find the forecasts.

$$\hat{Y}_n(l) = E(Y_{n+l} | Y_n, Y_{n-1}, \dots, Y_1)$$

$$= \theta_0 + \psi_l a_n + \psi_{l+1} a_{n-1} + \cdots$$

where,

$$E\left(a_{n+j} \middle| Y_n, \dots, Y_1\right) = \begin{cases} 0, & j > 0 \\ a_{n+j}, & j \leq 0 \end{cases}$$

Okay? So, as given before, the L-step ahead forecast, which is y n hat and then L within brackets, is nothing but the conditional expectation of y n plus L given all the information we have, which is nothing but y 1, y 2 up to y n. Okay? Now, a couple of very important points to note down here is we will discuss this equation first. So, what exactly is the conditional expectation of a particular error term having such a subscript? So, let us say n plus j given y1, y2 up to yn.

Now, you can pause the video here for a minute and then try to understand what is happening here. So, this conditional expectation would be 0 whenever j is bigger than 0. And why is that? Because whenever j is bigger than 0, This value would be what?

It would be either, let us say, a n plus 1, or a n plus 2, or a n plus 3, etc. So, the conditional expectation of any future error, because all these are future errors. So, what information do we already have? So, you already have information up to n. So, y1, y2 up to yn, which is precisely nothing but this set, right?

So, in order to find out any future conditional expectation of the error term, they have to be 0 because you do not have any information for that, right? At the same time, the conditional expectation of a n plus j would be a n plus j itself whenever j is less than or equal to 0. Now, again, think for a minute about what would happen if j is less than or equal to 0. So, if j is less than or equal to 0. These would mean something like a n minus 1, a n minus 2, a n minus 3, etc.

And, by the way, all the subscripts, be it n minus 1, n minus 2, or n minus 3, the information on these subscripts we already have, is it not? So, again, remember what information you already have. So, the information we already have is from 1, 2, 3, 4 up to n, right? So, the conditional expectation of any future error would be 0, right? While the conditional expectation of any of the errors where the subscript is between 1 to n would be the error itself. So, this is one very big assumption or very big fact that you have to kind of assume in order to write down this equation.

So, once you understand this conditional expectation in terms of errors, you then have to simply replace all the values here. For example, psi 1, For example, psi l e n, then psi l plus 1 e n minus 1, etc. Now, again, what happened to all the previous terms? So, you have to go back to the equation.

So, this was my combined equation. Now, note here that all these subscripts, right? For example, n plus l, n plus n minus 1, n plus l minus 2 are all beyond n, right? Now, again, coming back to the same conclusion that if the subscript is beyond n, its conditional expectation would be 0, right? So, what information are we kind of able to retain here?

So, we are kind of able to retain all the information or all the terms from a n onwards. Make sense? So, we will retain a n, we will retain a n minus 1, a n minus 2, etc. So, in the next equation, if you again see, these are the only terms that you have here and, of course, the constant which is theta naught. So, essentially, just to summarize this, the conditional expectation of Y n plus L happens to be nothing but this equation.

Now, what exactly is my forecast error? So, when you are forecasting something, there has to be some error in that, isn't it? And then that error is called a forecast error. And here, if you notice, I am using the notation involving E to denote the forecast error. And hence, I use a different notation to denote the errors by aps.

$$e_n(l) = Y_{n+l} - \hat{Y}_n(l)$$

$$= a_{n+l} + \psi_1 a_{n+l-1} + \psi_2 a_{n+l-2} + \cdots + \psi_{l-1} a_{n+1} = \sum_{i=0}^{l-1} \psi_i a_{n+l-i}$$

Okay. So, the forecast error is a kind of simple idea, which is nothing but the actual value minus the forecasted value. Okay. So, again, if you go back a couple of slides, you will One can write down the actual value, the equation of the actual value, which is y n plus l, and the equation of the forecasted value, which is y n hat within brackets l, and then subtract.

So, if you subtract these two, you will end up getting this equation. So, a n plus l plus psi 1 a n plus l minus 1 plus psi 2 a n plus l minus 2, etc., up to this last term. So, psi l minus 1 a n plus 1. And then this entire equation can be written down in a concise form, which involves a summation sign. So, summation i going from 0 to L minus 1 and then psi i and then a n plus L minus i. Now, my strong suggestion is that you pause the video if you want and then try deriving all these things on an individual basis.

So, the first thing is, is this summation exactly equal to the left-hand side, right? So, all these things you have to actually verify, right? So, the idea of the instructor is just to tell you the methodology or the steps, but then some of the things have to be verified on your end also. For example, is this summation equal to the entire left-hand expression that you see or not? By the way, a couple of properties regarding the forecast error are that the expectation of the forecast error, or in other words, the expected value of E and L, is 0, and the variance of the forecast error is nothing but the variance of this guy, which is nothing but sigma a squared, which is a constant, so it comes out.

$$\mathrm{V}\big(e_n(l)\big) = \sigma_a^2 \sum_{i=0}^{l-1} \psi_i^2$$

And in summation, i going from 0 to L minus 1, psi i squared. And why psi i squared? Because if you look at this expression, this contains psi i. So, if you bring out the constants, you have to square them. And obviously, the variance of all the error terms has been assumed to be sigma squared a, which is again a constant. So, these are some properties regarding the forecast error, which is E N L.

So, now what we will do is we will quickly revise some initial, let us say, one-step-ahead forecast or two-step-ahead forecast, etc. So, what happens if you have a one-step-ahead forecast? So, a one-step-ahead forecast means the value of L becomes 1. So, the problem is slightly simplified. So, this is the expression for the actual value Yn plus 1, and this is the expression for the forecasted value.

$$Y_{n+1} = \theta_0 + a_{n+1} + \psi_1 a_n + \psi_2 a_{n-1} + \cdots$$

$$\hat{Y}_n(1) = \theta_0 + \psi_1 a_n + \psi_2 a_{n-1} + \cdots$$

$$e_n(1) = Y_{n+1} - \hat{Y}_n(1) = a_{n+1}$$

$$V\big(e_n(1)\big) = \sigma_a^2$$

Now, again, the only thing you have to do here is to replace L with 1. So, wherever you have L in the previous slides, you replace the value of L to be 1, and then you get these two sets of equations, or in other words, the forecast error, which is En1 in this case, and nothing but the difference of these two. Which happens to be simply a n plus 1, alright. So, how much error are we making when we are forecasting just one step ahead is nothing but a n plus 1. Now, of course, under the ARMA model setting.

So, all these examples are for the ARMA PQ structure, or in other words, the variance of the forecast error is nothing but sigma square e. Similarly, what would be the two-step ahead forecast? So, this is the actual value of y n plus 2. This is the forecasted value of the two-step ahead forecast. If you take the difference of these two, it will give me the forecast error, right, which is nothing but a n plus 2 plus psi 1 a n plus 1.

$$Y_{n+2} = \theta_0 + a_{n+2} + \psi_1 a_{n+1} + \psi_2 a_n + \cdots$$

$$\hat{Y}_n(2) = \theta_0 + \psi_2 a_n + \psi_3 a_{n-1} + \cdots$$

$$e_n(2) = Y_{n+2} - \hat{Y}_n(2) = a_{n+2} + \psi_1 a_{n+1}$$

$$V\big(e_n(2)\big) = \sigma_a^2(1 + \psi_1^2)$$

And lastly, one can actually take the variance of the forecast error, which is sigma square a into 1 plus psi 1 squared. Now, obviously, in the subsequent slides, we will kind of deal more with some other ideas of forecasting and then not just let us say based on minimum MSE, etc. So, we will also talk about let us say some smoothing ideas or let us say Holt's method, Holt and Winter's approach, etc. Okay. Thank you.