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 25: Practical Session in R-5

Hello all, welcome to this course on time series modeling and forecasting in R. Now, again, if you see in front of you, we have the R window open. So, it means that in today's session, we will talk briefly about some practical angles of, let us say, forecasting. So, the entire session today will be focused on the practical aspects of forecasting. Now, again, if you remember very briefly, just to revise quickly over the last four sessions this week, we've talked briefly about different kinds of forecasting techniques.

So, of course, when the model is ready. So, again, by ready, I mean that you have fitted a particular model to the data set. Then you made sure that all the assumptions are being met or not. And then that exercise was called diagnostic checking. And once you're through with all these processes and all these stages, then the ultimate stage is forecasting, where you sort of forecast down the line in the future, depending on some historical information.

And again, if you remember, some of the forecasting techniques we've covered this week are some smoothing techniques. So, let's say simple moving average smoothing or exponential moving average smoothing. Right. And then we talked about double exponential or triple exponential, which had some other names also, such as, let's say, Holt's method or Holt and Winter's approach, and so on and so forth. Right.

So, particularly in today's session, we'll tie all these things down and then try to explain, using a practical data set, how one can forecast down the line if you have some available historical data with you. Now, again, this would be the typical R window to start with. And then again, we'll try to import some libraries. So, the first library we want is called TTR. So, as you see in front of you in line one, the full form of TTR is interesting.

So, it's called technical trading rules. Right. So, if you remember, if you want to forecast any stock price or something like that, people often apply some technical indicators. So, let's say moving average, RSI, or MACD. So, these are some well-known technical indicators.

So, this package helps us to draw some of the moving averages, implement them on some real data sets, and eventually forecast. All right. And again, if you want to run a particular line on Windows, you can either click on this run that you see here or press control enter. Similarly, if you have a Mac and you're working on a Mac, you can directly press command enter to run a particular line. All right.

Okay, now initially in this video or in this session, we will focus on two specific data sets, and probably we have covered these data sets in the earlier code sessions also. So, the first one is called the NOTM data set. So, we will fetch the data first. Let us say data and then NOTM. So, the name of the data set is NOTM, right?

Now, again, if somebody does not remember what this data set means, then I can quickly write down something like a question mark and then NOTM. So, if you do something like a question mark and then NOTM and then hit enter, it sort of gives you a small description of what exactly the data set is. So, the NOTAM data set focuses on this thing. So, the average monthly temperatures at Nottingham, and by the way, Nottingham is a county in the UK between all these years. So, let us say 1920 to 1939, right?

Now, a couple of things to focus on here. So, you have monthly temperature data, and immediately if you have any monthly temperature data, then the immediate thing that should come to mind is that there should be some seasonality present, right? Because we discussed this a number of times in the previous lectures also, right? If you are working with any temperature data per se, so temperature data, or let us say rainfall data, humidity data, all these contain some seasonality aspects. And then if you go down the line, it sort of gives you a small description also. So, a time series object containing the average air temperatures at this particular location, which is Nottingham, and particularly Nottingham Castle.

So, you have this Nottingham Castle in the county of Nottingham, or the town of Nottingham in the UK, and then in degrees Fahrenheit over 20 years. So, 1920 to 1939, right? And, by the way, this is the source. So again, if somebody is more interested in delving deep into a particular data set or a particular function in R, then R nicely gives you the source of the data set also. So, this data set was picked, or it was kind of discussed first in this book here by Anderson.

So, time series analysis and forecasting, the Box-Jenkins approach. All right. And then, down the line, it gives you some examples also. So, let's say if you want to implement this data set using some basic techniques, right? So, how can you do it, right?

So, getting the help command by using the question mark sign. So, a general syntax is a question mark followed by any name of the data or the name of the function that you want some information on. And R nicely combines several details about the particular data set or the function from which you can get multiple further ideas. So, coming back to this, we will work with this data. So, we fetch this data, data NOTM, and then force NOTM.

So, force NOTM just ensures that you are kind of sticking that data inside the R environment, nothing more than that. So, let us say force NOTM, and on top of that, you can see that it sort of gives you the actual data also. So, let me scroll up, and then again, since I told you that you have a monthly sort of data. So, January, February, March, April, all the way up to, let us say, December, and then over all the years. So, the first column kind of gives you the years, and then all the subsequent columns give you the months, and all these values that you see inside the table are nothing but the actual temperatures in Fahrenheit.

All right, now the essence of kind of using this data, and then the first thing that we do is sort of decompose, right? And then we've talked about time series decomposition sort of long back in one of the earlier sessions. So, time series decomposition, if you remember, is nothing but to split the dataset into its several parts. So, let's say trend, seasonality, or irregular fluctuations, etc., okay? And in R, you have this very easy command called, let us say, STL.

So, STL sort of splits the data into different components. So, it says here that decompose the time series into seasonal, trend, and residual factors. So, we will see what this command gives you. And then we have given it a name. So, NOTM underscore STL.

Now, what we will do next is plot the actual data. So, plot NOTM right, and then here you see the actual plot of the data. So, I will sort of zoom this in a short while, right? And then the second plot we will do is NOTM_STL, which is the decomposition plot, right? And this is a very typical decomposition plot. So, what we will do is we will sort of zoom this plot, right? Because the decomposition plot also gives you the underlying data, right? And then this is a very typical time series decomposition, right?

And then probably we have definitely seen this sometime in the previous lectures also. So what it tells you is how the actual data sort of behaves in the first row. And in the second row, it sort of gives you the seasonal aspects of the time series as to how the seasonality angle behaves or how the seasonality aspect behaves, right? And then by looking at the second row, one can immediately tell that there has to be some underlying significant seasonality, right? Because you see some constant peaks and constant drops there, right?

Then the third row tells you if the data set contains any trend or not. So, what do you think? So, do you think that the data set contains any trend? No. Probably not, because otherwise you will see a clear trend here.

So let us say upwards or downwards, right? But then, since this is sort of in between some fixed horizontal bands, one should not expect that there should be a very high trend in the data or something like that, right? And then again, the last row sort of tells you the remainder aspect. So once you remove the seasonality and the trend, whatever remains in the data is nothing but the remainder. So my strong suggestion is that if you're trying to sort of analyze or even go ahead and do forecasting using any practical data set, all these are sort of initial ingredients that one should focus on.

So let's say time series decomposition, some initial plots of the data. So let's say a scatter plot or, you know, all these are kind of starting points. If you want to analyze a practical data set, then eventually the last step would be forecasting. Of course. All right.

Okay. So we'll close this now and then go back to the code. Now, using the decompose series, we want to extract some time series. So, let us say 1 to 12 and then the first column. So, what do you get by doing this is nothing but the data set, right?

So, this is sort of telling you some adjusted series, right? Because here you see some negative values, right. So, if you want some adjusted series, right, then you can sort of use such a command, basically, okay. All right. So now we'll shift to another data, which is sort of air passengers. And then again, I'm pretty sure that we've used this air passengers data quite often in the previous code sessions or previous practical sessions.

So we shall first fetch the air passengers data. And again, the same thing. So again, if you don't remember what this data set is all about, then probably again, you can look at the earlier videos or use the technique where you click the question mark and then write down the name of the data set, which is air passengers. Now again, one thing here, if you notice, is that you need not even write down the full name of any function or any data set which has been called in the R environment. So since we have already called this data, you can write down 'air' and then basically click one of these.

So I will click the first one because we want some information on the actual air passengers data set. Right. And then you sort of hit run or, you know, and then again the same thing. So if you look at this window here, it tells you a small description about the air passengers data. Now, again, I'm pretty sure that many of you might remember this data set.

So, we work with this dataset. And again, the reason why I picked the same dataset is just to obtain some consistency over what we are doing in the practical sessions, right? So, if you bring in some random datasets every time, then you might not be able to connect the lines there, right? So, since we worked with this air passengers data even before, let us say for model fitting or diagnostic checking. So, why not just extend that and then try some forecasting techniques, okay?

But again, just to quickly revise what this dataset means. So, this air passengers data gives you the monthly airline passenger numbers between 1949 and 1960. Okay. And again, a very small description. So, let us say the classic Box and Jenkins airline data.

And again, the same thing. So, monthly totals of international airline passengers between 1949 and 1960. Okay. And then, how would you use it? So, usage, right?

And then format. So, format is, this is a monthly time series in thousands, okay? And lastly, this is the source. So, Box and Jenkins and Reinsel. And then this is again a book.

So, time series analysis, forecasting, and control. And then some examples. So, let us say immediately if you want to run some ideas on this dataset using the air passengers data, then using some of these examples, you can actually, let us say, fit some plots or, you know, fit some linear regressions probably and then test some things out and so on and so forth. Okay. Okay.

So, again, we have already fetched the data using the data air passengers command. Now, again, what we will do is we will again decompose. Right. So, again, like I said, time series decomposition has to be there if you are sort of analyzing any practical dataset as a very initial step. Okay.

So, and then here, if you notice, we will use this command instead. So, decompose. So, lastly, for the NOTAM dataset, we use the STL command. But then one can do it using either STL or decompose. And here one can actually specify whether we want an additive decomposition or a multiplicative one.

So, all these things we have kind of discussed in the theory session. So, additive means that each component—trend, seasonality, and irregular components—are sort of added together. So, T plus S plus I, as opposed to that, if you have a multiplicative sort of type, then we sort of multiply each of the components. So, trend into seasonality into I. So, we will see how the decomposition looks like, and then now what we want to do is we want to extract the seasonal components out of that.

So, let us say if you want to extract the seasonality aspect of the dataset, then this is exactly how the seasonal component would be fluctuating over all the months. So again, the first column gives you the years, and then each subsequent column gives you the corresponding months, and then all these values sort of tell you the seasonality component for that particular year or that particular month. Now here, clearly you can see that the seasonality values do not change over the years because what is happening here is, let us say, minus 24.74. So, this value is the seasonal value, and right, in the month of January, or let us say, minus 36.18. So, this is the seasonal aspect of the time series in the month of February, and so on, right.

And obviously, as you go down the line, let us say July, August, so 63.83, 62.82. So, all these are seasonality values extracted from the decomposition for that particular month, all right. So, such things could be done once you sort of have a decomposition of a time series and so on and so forth. And now, if you want to see that plot, right, if you want to obtain that decomposition plot, you again have to apply the STL technique. So, STL_air is nothing but STL applied on air passengers.

And then S.window is periodic. So, this is more like a syntax. Okay. All right. So, we see how the decomposition looks like.

And again, let's say if you want to extract some information from the decomposition that we have done, then one can actually do it using this way. Now, again, if you want the plot, right, so if you want the plot of the decomposition, then I can actually scroll up, and you can actually use plot and then the decomposition in terms of STL. So, here, since we have given it a name of STL_air, so I can simply do something like, let us say, plot and then inside round brackets, I can write down STL and then underscore air. Okay, because this was the name for the decomposition that we have done, right? And then, if you run this, then you will see that you will get the decomposition plot here, okay. So, let us zoom this and then see what is going on.

So, again, the same story. So, the first row gives you the actual air passenger data and then how it behaves, and now the second row sort of tells you the seasonality aspect of that, and then clearly again, you can see or you can tell that there is some substantial, significant seasonality which is present, right? And now, surprisingly or interestingly, you also have a trend, right? Because the third row sort of indicates very clearly that you have a very strong upward-sloping trend. And then, lastly, the remainder aspect.

And here, one more interesting aspect in the remainder part is that can you see some periodicity in the remainder? Right. So, the remainder itself is not completely random. Right. And one more thing is that if you go down the line here, let us say 1958 or 1959 or 1960, then you can clearly see that the variance is also increasing.

Right. Initially, all the spikes are kind of very tight. But as you go down the years, let us say towards the end of the time frame, you can actually see that all the spikes are kind of expanding, and then the variance is actually increasing. Also. So, air passenger data is a very famous data which people analyze, and why?

Because this is completely non-stationary data to start with, and this data sort of contains all three aspects of non-stationarity. So, trend, seasonality, and changing variance. So, the trend is there, seasonality is there, and the changing variance problem is also there. So, how do you analyze such data? So, how do you model it?

Then, how do you apply some diagnostic checks, and eventually, how do you forecast? This is the kind of theme for many experimenters or analysts. Okay, so we will close this dataset and then go back to the code. Now, here, what we will do is try some smoothing techniques. So, eventually, our goal is to do some forecasting or apply some forecasting techniques. So, why not just start with some basic smoothing techniques?

So, for that, what we will do is have a simple plot of the dataset. So, this is exactly how the air passengers data looks. And we have seen that in the previous decomposition plot as well. Now, this command that you see in line 29 just tells you or ensures that you are considering this air passengers term or the air passengers data as a time series object. So, we added this TS in front of that.

So TS applied to air passengers tells you that you have to actually consider this as a time series object. Nothing more. Now here, what we are doing is we are fitting a simple moving average smoothing technique. So can you see that? So SMA stands for simple moving average.

And what is the order of that? So again, if you go back, let's say a couple of sessions this week. Right. Then you should always remember that any SMA or any exponential moving average has some order. Right.

And then we've discussed that also. So can you. So the idea is to play around with the order. So can you sort of change the order? And then the idea is that if you change the order, which order sort of gives you the best fit?

This should be the ultimate goal, shouldn't it? So, I can play around with the order. So, let us say I can take 3, 4, 5, or let us say 7, 8, and so on, and I can see how the SMA is fitting the actual model or how the smoothing technique is being applied to the actual dataset. Then, from all the visual plots and visual checks, I can pick the order where I think the fit is better or the best. So, here initially, what we are trying to do is we are fixing the order to be 3, right? So, again, air passengers underscore TS underscore a simple moving average smoothing technique where the order is 3, right?

And now we will see what happens. So, if you run this, then this is now stored in this name, right? And now what we will do is we will see the actual data. So, this is the actual data, okay? So, this is the actual data, right?

And then the next thing we will do is we will actually see the SMA values. So, since we fitted this SMA 3 model, right? So, now we will see what values the moving average or the smoothing technique with an order of 3 is giving us, right? So, I think this is just for comparison. So, we can place these two datasets side by side and then probably gauge whether they are giving us a

Superior model fit, sort of, or is this SMA 3 indeed a better fit? Right? And how could you do it? I mean, you can compare, let us say, piecewise. So, let us say the first value is 112, then 118, 132, 129. Now, remember one thing: since you are applying a simple moving average of order 3. Now, again, if you go down, so these are the moving average values. You will see NA for the first and the second one. And why is that? Because there won't be any values for the first and the second one since you are using a window size of 3.

Because the idea behind SMA is that you go back in the history and then you are taking the average of the last 3 values. So, since you are using SMA 3, then there has to be no value here in the first one as well as no value in the second position. Now, the third value is 120.66. Then the fourth value is 126.33. Now, you have to actually find out: are they matching with the actual observation?

So, here the third value is 132. Then the next value is 129. So, this is just a rough idea as to how you can compare by keeping these two subsets of the data side by side. So, the actual data and the SMA3 values. But one can actually plot the SMA3 values fit also.

So this is nothing but the simple moving average smoothing technique, and then the plot of that that you see here. And the command that you see in line 34 is essentially what would happen if you subtract the two. So rather than comparing one data point with the other from the two subsets, why cannot we simply subtract and then see if the subtractions or the resultant subset is close to 0 or not. So if all the values are sort of close to 0, then we will say that piecewise they should be matching. And then here what we are doing is we are subtracting the actual data and the SME3 values and then plotting that.

So, this is that plot. Let me zoom this right, and then we sort of discuss what is happening here. So, firstly, if you see that pretty much the values are kind of revolving around the line 0, which is exactly what we want, right? Because if you take the actual data and subtract the fitted data. So, something like y minus y hat, similar to what we do in regression also, right? So, these are nothing but the residuals in some sense, right? So, the residual should revolve around 0 because all the errors should be really close to 0.

But there might be some problem here because you see some periodicity. So you have like peaks and troughs which are kind of repetitive. And then down the line, you can clearly see that there is a changing variance problem also. So as you go down the line, then the variance aspect in the underlying data set or the underlying residuals is sort of also increasing. So these are some observations that the experimenter can make by simply plotting the residuals or nothing but y minus the y hat values.

Alright, now what? So, now as a comparison, what we do is we will plot or find out another SMA with an order of 8, right? And such things have been done very easily by changing this n to whatever number you want. So, this n is the order of the SMA. So, earlier if you go back, then here the n was 3, but then here if you see, I am changing the order from 3 to 8.

Now, how will the performance be? So, we will find out. So, let us say we found out the SMA 8 values. This is the actual data. This is the SMA 8 data, right?

Now, this is exactly how the SMA 8 plot looks like, and this is exactly how the plot of the residuals looks like. So, again, let me zoom this one, right? And then if you zoom it further, then this is exactly the behavior of the residuals of an SMA 8 fit, right? And then one problem here I will say is that if you compare this plot with the earlier residual plot where we used SMA 3, one can clearly see that the amount of changing variance is predominantly high in this fit because all the peaks, if you see, are really higher than the earlier plot. This could be a problem because we want the variance to be as less fluctuating as possible.

So, if the variance is fluctuating that much in SMA 8 as compared to SMA 3, then probably one should actually go with SMA 3 instead. So, these are just again some observations that one can actually make from creating such plots. So, this was SMA, and now EMA. So, EMA is exponential moving average smoothing. So, again, the same thing.

So, we will try to fit two EMAs. So, the first one would be of order 3, and then the second one would be of order 8. So, let me run both of them simultaneously. So, the first one has an order of 3. And then the second one, which is given by this name.

So, EMA 8 underscore air passengers is with an order of 8. Now, again, the same thing. So, we will create three plots. The first one is the EMA 3 plot, right? This is exactly how the EMA 3 smoothing fit looks like.

Then the second one would be EMA 8, right? So, this is exactly how the EMA 8 smoothing fits, right? And then the third one is of an order of 12 now, right? So, if you see here, n equals 12. So, how would that plot look?

So, this would be that plot, right? Now, one feature or one observation which one can find out using all three plots is the EMA 3 plot, then the EMA 8 plot, and then the EMA 12 plot. So, if you go on increasing the order, can you see that you are getting smoother and smoother plots towards the end? So, this is EMA 12. Let me show you EMA 8.

EMA 8 is this, and then EMA 3 is something like that. So, can you see that EMA 3 would be more suitable for the underlying dataset as compared to, let us say, something like EMA 8 or EMA 12, which are much smoother? The idea is that we want to preserve the underlying tendencies of the dataset also. So, if you have a really smooth kind of EMA or SMA plot, then that might not serve the purpose. So, now we are almost there.

So, now the next thing we will do is we will try to apply the Holt technique. So, Holt exponential smoothing, which is also called double exponential smoothing. Now again, if you remember, we discussed in the last lecture that if you want to implement this in R, you have this common command called HoltWinters. So inside this HoltWinters itself, you can actually specify whether you want a Holt technique or a Holt-Winters approach. Alright, so we will run the HoltWinters command, and here, how do you specify that you want a Holt technique and not a Holt-Winters technique?

By using this syntax. So, gamma equal to false. If you specify gamma equal to true, then it also estimates the gamma parameter, and then the gamma parameter is associated with Holt-Winters and not the Holt technique. So if you specify gamma to be false, this is actually the Holt technique. Now, again, the same thing.

So, this is the value. So, by the way, this is what the syntax of the output of the Holt technique tells you. So, it sort of gives you the optimum alpha value, which is 1, and the optimum beta value, which is 0.0032. And then the underlying coefficients of the Holt technique. And now, can you extract the SSE value of this fit?

So the SSE value happens to be that much. So 163634.1. And then eventually, can you forecast it or can you plot it? So this is nothing but the Holt-Winters filtering. And then this is, by the way, simply Holt and not Holt-Winters.

And then you can see that I am kind of getting a very superior model fit. And now, if you want to forecast in the future, so for that, you require this library which is called forecast. And then what we'll do is we'll sort of forecast the air passengers data in the future. So we'll use this command called forecast. And then we'll sort of give you a plot of the forecast also.

So can you see here? So this is the actual data. And this is exactly how the forecasts are looking. And now you might wonder, which technique is the forecast in? So again, if you come back to the code, since we specified air passengers F, so air passengers F is again, if you scroll back slightly, this is given by the Holt technique.

Because this is the name we gave to the Holtz technique, right? Air passengers F. So whatever fit you specify here, in this case, is the Holtz technique, then the forecast would be based on that or using that technique. And here you can clearly see that the forecast is not matching the actual dataset, right? I mean, they are not able to preserve the seasonality. Can you see that?

Because the forecast is almost a horizontal line. It is not preserving the seasonality aspect. So, can you do better? The answer is yes. So, now the last thing is we will try to implement the Holt-Winters approach because remember one thing: if the underlying dataset contains seasonality, there is no point in applying the Holtz technique.

One should actually go with Holt-Winters. So, we will do the exact same thing, but then the small tweak here is that we are applying Holt-Winters on the log of the data. And why? Because again, if you remember, there was this changing variance problem. Now, how do you correct the changing variance problem? By applying some transformation.

So, let us say log or square root, etc., right. So, we will apply this Holt-Winters technique on the transformed data set, which is log, all right. So, this is the forecasting or this is the model fit, right. And then this is the SSE value, 0.2030, and this is the Holt-Winters filtering. Again, you can see that it sort of preserves the seasonality and the trend, etc., right.

And the last thing is forecasting. So, forecasting using which method now? So, forecasting using the Holt-Winters approach. And here you can clearly see that if you are actually applying a Holt-Winters approach rather than Holt's technique, you are actually preserving both the trend and the seasonality, etc., right? So, the summary here is that if you have both trend and seasonality, one should actually go with the Holt-Winters approach so as to preserve the trend aspect and the seasonality aspect rather than something like, let us say, SMA or EMA or Holt's technique.

Thank you.