Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 03

Lecture 13: Cyclicality and Test for Stationarity

Hello all. So, welcome to this new video in this course on time series forecasting and modeling using R. Just to give you a short background on what happened in the last session, we were talking more about a few of the plots which try to measure seasonality. So, we started with a few examples showing what exactly you mean by seasonality, and then towards the end, we saw some plots capturing seasonality visually. And there, if you noticed, we talked about all these kinds of plots. So, the first one was, let us say, a seasonal plot, right? And then the second one was called a seasonal sub-series plot, right? etc.

Okay. And then each plot has its own advantages and disadvantages. So, probably initially in today's session, we will talk about one more very interesting plot, and then we can move on. So, I thought maybe I could draw the plot rather than showing you the actual plot. So, this plot is called individual box plots as per seasons.

So, what we have is that again we have an x-axis and a y-axis. And then, let us say, again on the x-axis, we have the months, all right, something like that. And then on the y-axis, we have the actual time series values. Now, typically, as we have seen in the previous plots also, be it a seasonal plot or a seasonal sub-series plot, right, etc. On the x-axis, what we had were months, right.

So, let us say if you have monthly data, something like that. So, January, February, and then March, all the way up to December. So, the last one is December, right? Now, how exactly did you create the seasonal plot? So, if you remember, the seasonal plot is nothing but for each particular year we have values like that.

So, let us say in one year one can see a pattern like that. In the next year, one can actually see a pattern like that. So, one can actually draw the individual patterns corresponding to

each year over all the months. So, this is a typical seasonal plot. So, if you erase this, then the next one was called a seasonal sub-series plot.

So, there, what we had, if you remember, is in a particular month we had individual line plots for that particular month, right, something like that. So, these line plots kind of tell you the extent of the data in that month over all the years, right? So, for example, January. So, in January, what happens is, let us say this is the minimum value across all the years in January, and this is possibly the maximum value across all the years in January. So, if you remember the airline passengers' data, the data ranges from 1949 to 1960, typically, right? And then, during January, if you look at

The number of airline passengers who traveled in January in each of the years from 1949 to 1960, this is the extent of that line plot. And a seasonal sub-series plot is one more slightly added advantage is that one can actually plot the mean also using a blue line. So, let us say if you draw a blue line something like that here in between somewhere. So, this horizontal shorter line kind of captures the mean. So this is a very typical seasonal sub-series plot.

So now what we will do is we will try to understand one more plot which is called as individual box plots as per seasons. So individual box plots as per seasons. So the idea is kind of very clear here is that given any month I can actually create some box plots for that particular month. So let us say this would be my box plot for the month of January.

This might be the box plot for the month of February. This might be the box plot for the month of March, etc. And then towards the end, similarly, this might be the box plot for the month of December. Now, again why is such a plot again important and then valuable is that box plot in itself tells you a lot of things right. So, the extent of this box plot.

So, these are called whiskers of the box plot, by the way, and then you have the actual box, right. So, the whiskers kind of extend to, let us say, either the minimum or the maximum of the data, something like that. And in the box, the box itself kind of extends from Q1. So, Q1 is the first quartile, all the way up to Q3. So, Q3 is the third quartile, right?

And then, in between all these boxes, we can actually create the medians also because Q2 is nothing but the median, which is given by all these horizontal lines. So, let us say this is exactly where my Q2 lies. But then, Q2 is nothing but the median, okay? So, a box plot in itself tells you a lot of stories about how the data is spread, or where exactly the

quartiles are lying, or where exactly the median is lying. And on top of that, one can actually again get an idea about the extent of all the values for a particular month over all the years.

So, I will say that this kind of a graph, or this kind of a side-by-side box plot for each month over different years, is kind of similar to a seasonal sub-series plot. The only difference is, in a seasonal sub-series plot, you draw a typical line along with the average, something like that. But then, here we are actually incorporating much more information by drawing individual box plots. So, again, the idea is that one can actually pick and choose to their liking as to which plot should one draw, but then the more information a particular plot is kind of able to convey, the better it is. Alright, so now we will kind of shift attention from seasonality to another aspect of capturing non-stationarity, which is called cyclical variations or cyclicality, right.

And then, in a short time, we will kind of differentiate between seasonality and cyclicality. So, even though they may seem similar, seasonal variations and cyclical variations, you do have some subtle differences, alright. So, firstly, what exactly do you mean by cyclical variation? So, cyclical variations are gradual, relatively long-term, and then up and down kind of irregular repetitive movements of a time series, alright. So, this definition has a lot of individual pointers.

So, let us say cyclical variations are gradual, relatively long-term, long-term, then seasonal, and then up and down, irregular and repetitive movements of a time series, ok. The second point is that the data shows rise and falls which are not of a fixed frequency. So again, if you remember when we talked about seasonality, all the variations and all the fluctuations were of fixed frequency, right? So again, if you remember a typical temperature plot. So, how does a typical temperature plot look like?

So, it could be something like that, right? So, all these peaks and all these troughs are repeating themselves at a fixed frequency. So, the difference between any two peaks or the difference between any two troughs is kind of similar, roughly, right? But when it comes to cyclical variation, this is not the case, right? So, the distance between the first peak and the next peak or the first trough and the next trough may be irregular, all right?

And the third point is that the period usually extends beyond a single year. So, again, if you remember in seasonality, we had the period to be less than one year, right? So, again, in temperature data, one can actually see repetitions that are happening in the same

months in different years. So, the period does not extend beyond one year, right? But when it comes to cyclicality, the period usually extends beyond a single year.

And then you have like six phases in any cyclical variation as follows. So, expansion, peak, recession, depression, trough, and recovery. So, any cyclical variation or any cyclical pattern over several years, let us say 10 years, 15 years, 20 years. Would be comprising all these six stages or all these six phases. So, expansion phase, peak phase, recession phase, depression phase, trough phase, and then recovery phase.

Now, again, probably in the next slide, we will kind of elaborate a bit more on all these phases and then what exactly do you mean by all these individual phases. Alright, so the graph that you see in front of you kind of beautifully demonstrates the six phases of a business cycle. So, typically, one can observe cyclicality or cyclical variations in a business cycle. So, let us say you have a business. So, the business might be in a boom or the business might not do that well, right?

So, all these kinds of indicators let us say peaks or troughs or periods of recession or periods of expansion, etc. So initially, let us say the business cycle is expanding, and then it expands to the peak. So, when the maximum is attained, this is the peak, and then from the peak, when it kind of comes back to the zero line, it is called a recession. So, recession means that the business cycle is not performing that way. So, it is going on a downward kind of path or trajectory.

And then if it again continues to kind of behave in that same manner, going down in a downward trajectory, then it is called a depression, and then the trough is attained. And then, once a trough is attained in a business cycle, it again starts to recover. So, the recovery phase begins, and then the same thing repeats. So, expansion, peak, recession, depression, trough, and recovery. And once this recovery stage is attained, now again the second cycle would begin.

So, this is a typical cyclical variation of a business cycle. Now, let us look at a few examples, and then we will try to differentiate between cyclicality and seasonality. So, let us say the first example could be business cycles. So, you talk enough about business cycles. So, let us say economic expansions are followed by recessions.

So, typically something like this. So, if a business cycle is expanding, it attains its peak, and then the recession starts. So, something like that, right? Let us say this is the horizontal line, and then once the recession has kind of reached the zero stage, then again

what you can do is you can come down. So, let us say it attains a trough and then again comes back, something like that, okay. So, this is a very typical business cycle.

Then one more example could be the price cycle. So, prices of let us say some commodities might as well behave like that. So, let us say the prices are expanding, then the prices reach their peak, right. Then once the prices reach their peak, they again start a downward trajectory. So, this is called recession, then depression, then trough, and then recovery.

So, when future production decisions are based on current prices. So, let us say based on current prices, if you want to kind of predict what would happen in the future in terms of production, then this is a very typical price cycle, or one can actually observe some cyclical variations in natural phenomena. So, let us say solar cycles because what happens is the sun's magnetic field undergoes a cycle after approximately every 11 years. So, the magnetic field of the sun itself kind of behaves in a cyclical manner and then sort of repeats after every 11 years. So, after every 11 years, what typically happens is that the north and south poles of the sun kind of exchange their positions.

So, they shift entirely. So, it's kind of a similar idea to a business cycle or a price cycle. So let's say when the North Pole is towards the north or in one direction, then the stage could be an expansion stage. And then probably when it starts flipping, then we can actually attain a kind of recession stage, then probably a depression stage, and then again, a recovery stage. So, all these are a few examples of where one can observe cyclicality or cyclical behavior in, let us say, natural phenomena or price data or, let us say, business cycles and so on.

So, now talking about the sunspot data. So, very typical sunspot data. So, what you observe here is that Let us say this is the behavior of sunspots. So if you observe on the surface of the sun, you have these dark spots, right?

And then those dark spots are called sunspots. And then ideally, understanding or analyzing the number of sunspots appearing on a daily basis or monthly basis, right? It is kind of important because that decides the magnetic field of the sun, right? In any case, let us say this is exactly how the sunspot data behaves. So, the yellow line is the daily observations, the blue line is the monthly data, and then the red line and possibly towards the end here are the forecasts and the fitted models.

All right. So how exactly does a particular fitted model behave? But then the actual data is given by yellow and then blue. Right. And then even here, you can see the same typical behavior.

Right. So let us say the number of sunspots is growing here. Then they start falling. Right. And then this might be a recovery phase where it starts to pick up again and then possibly it will come down again.

So again, the number of sunspots behaves in a cyclical manner. And then here, I'm using some images which are due to all these kinds of citations. So, if one is really interested in what exactly is meant by sunspots, how are sunspots measured, and what is the use case of doing all that? So, one can actually refer to all these links, by the way. Alright, now the important pressing question is what exactly is the difference between seasonality and cyclicality?

Right, because again, in seasonality, we have repetitions, and in cyclicality, we also have repetitions. So, what might be the difference? Okay. So, if the fluctuations are not fixed or they are not of a fixed frequency, then you have a cyclical variation. So, it’s something like this.

So let us say if you draw a random graph and if the repetitions are not of a fixed frequency, so let us say the next one could be something like that, then possibly the next one is like that. So even though you have peaks which are here, here, and there, but then those are not of a fixed frequency. So, this is a very typical graph of a cyclical variation pattern. On the other hand, if the frequency is unchanging or unchanging means fixed. So the frequency is unchanging and related to some calendar effect; it is called seasonality.

So temperature data, rainfall data, data due to holidays, so sales during Diwali, etc. So, all these examples are seasonal examples. And typically, the average length of cycles is much longer than the seasonal patterns. We have seen that, right? Because a typical business cycle does not end within a year, right?

So, it might take, let us say, 10 years, 15 years, or 20 years to complete one business cycle comprising all the six phases that we saw earlier, right? So, the average length of cycles is much longer than the typical average length of a seasonal pattern, right? And last one is, the magnitudes of cycles are more variable than the seasonal patterns. So, the amount of variation in the peaks or the amount of variation in the troughs is much more

variable in a cyclical pattern than in a seasonal pattern. So, these are some of the differences when it comes to understanding the ideas behind seasonality and cyclicality.

Because even though these two are non-stationary aspects, they mean entirely different things. So, how do you measure seasonality? How do you measure and capture cyclicality? So, you have to kind of come up with slightly different ideas. So, probably what we will do now is shift attention to a slightly different kind of angle, and then we are almost there towards defining that SARIMA model and then understanding a few more aspects about non-stationarity.

So, firstly, we will try to understand, as in time series literature, we have a thing called a unit root. So, what exactly do you mean by a unit root? So, let us say, consider a model. So, maybe you can pause the video for a minute or something and then try to identify this model. So, we have discussed this model a number of times.

Okay. So, this model is a typical ARMA model, isn't it? Because this is the autoregressive structure. This is the moving average structure. Okay.

So, this model could be a typical ARMA kind of model with what orders? So, since it goes back to t minus 2. So, the autoregressive order could be 2, and then the moving average order could be 1. So, this is an example of an ARMA 2 1 model. Okay.

Now, if you focus on the left-hand side of this equation of the above equation. So, what do you observe? So, I can actually write down the equation in this polynomial structure. Now, again, what exactly is B here? So, B is nothing but the backshift operator, all right.

So, B is nothing but the backshift operator. So, if you operate B on y t, this gives you y t minus 1, all right. So, B is the typical backshift operator which we studied in one of the previous lectures, all right. So, again, if you focus more on the left-hand side, this is exactly how you can rewrite the equation, is it not? Because I can actually combine the coefficients. So, 1 minus 1.9 into b plus 0.9 into b square applied on y t. So, this entire thing is more like an operator applied on y t, all right.

$$Y_t - 1.9Y_{t-1} + 0.9Y_{t-2} = e_t - 0.5e_{t-1}$$

The LHS can be written down as:

$$(1 - 1.9B + 0.9B^2)Y_t$$

Thus, the roots of $1 - 1.9B + 0.9B^2 = 0$

Now, if you decide to solve for such an equation, you take this operator as it is. So, 1 minus 1.9 b plus 0.9 b square equals 0. So, it turns out that the roots of this equation are of interest to us. So, what exactly are the roots of this equation? So, one can use any calculator or try to solve this using pen and paper and then find out the answer.

So, the roots of this equation are 2 because you have a quadratic equation. So, there should be two roots. So, the roots are 1 and 10 by 9. And here, if you notice, one of the roots is exactly equal to 1. So, whenever such a thing happens, we call that root a unit root.

So, we hence have a unit root. So, whenever one of these roots is 1, or whenever one of the many roots—because this can be extended up to many more lags, right, depending on this autoregressive order, right. So, if one of the roots of many such roots of this typical polynomial happens to be 1. So, that root is called a unit root, right. Now, why exactly is studying unit roots important and so on?

So, we will study it in the next slide. Okay. So, firstly, unit roots make a process non-stationary, but then exactly why? Right. So, however, if you assume, let us say, Wt, which is a difference process.

$W_t = \nabla Y_t$ we get,

$$w_t - 0.9w_{t-1} = e_t - 0.5e_{t-1}$$

Okay. So, unit roots make a process non-stationary to start with. So, whenever you observe a unit root as one of the roots of that equation, you can actually say that the process is not stationary, right. But why exactly? So, let us say the initial process yt is non-stationary.

So, y t has to have one unit root, ok. However, if you assume the differenced process, so let us say w t is what? So, w t is nabla operated on y t. So, you are differencing it once, ok. So, essentially what we are doing is we are getting this. So, again if you go back one slide,

So we started with this model right here in terms of y t. And here if you apply difference in 1, so if you kind of calculate something like that, which is nabla y t, which is of course w t as per our notations, then eventually you end up getting this equation. And then this is nothing but a stationary ARMA 1-1 process. Make sense? And hence we argued the initial process y t follows an ARIMA process with 1, 1, 1.

So, initial process y t happens to be ARIMA process 1 comma 1 comma 1. And why exactly? Because you are differencing it once and you are getting hold of a stationary ARMA 1 1 process. So, the original process has to be ARIMA 1 comma 1 comma 1. And since in the original process you had a unit root, so unit roots kind of make a process non-stationary.

Because the ARIMA process is an example of a non-stationary process to start with. Okay. All right, now the end story is, why is a unit root a problem, right? So, what one can do is, unit roots kind of imply the following model structure. So, whenever you encounter a unit root,

I can actually write down this equation. And what exactly is the equation? So, 1 minus b, where b is the backshift operator applied on yt, right? So, if you expand this, what will you have? This is nothing but yt minus yt minus 1 equals some function of white noise terms or some function of error terms, right?

So, et's are nothing but errors, right? In our notations, So, let us say if yt minus yt minus 1 is a pure function comprising of some errors. So, f of et is what? f of et is some function of error terms. This implies that the current value, which is yt, is the same as the past value plus some white noise term or some white noise structure. Because yt would be what?

So, yt would be yt minus 1 if you take this on the other side plus a function of errors. So, what do you mean by that? Thus, in the long run, the process will never converge to any mean, right? And hence, this means that the process is, in fact, non-stationary. So, in the long run, if the process does not converge to a particular mean, right, we can actually say that the mean is changing or the mean kind of depends on the time, right? So, under all these assumptions, we can actually say that the process itself is not stationary, ok? So, whenever you have a unit root in that polynomial structure or in that polynomial equation. If you observe at least one unit root, then the process has to be non-stationary.

Because of this simple reasoning. Because the current value depends on one past value through some errors. So, the current value is the past value plus some random errors. Right. And then, what exactly those errors would be? Nobody knows.

Right. So, the past value plus a bunch of errors is giving me the current value. Right. So, in the longer run, it would never converge to the mean because the entire process depends on those random errors, ok. And hence, the unit root is a problem, ok.

So, now, we are at a stage where one can actually define the causes of non-stationarity, right. So, any practical time series can be non-stationary because of the following reasons. So, the first one is the presence of unit roots. So, we argued a short while back as to why that is a problem. Then, the next one is the presence of a deterministic polynomial trend.

So, what does a deterministic polynomial trend mean? So, the time series has some trend, and that trend aspect is deterministic. You can actually determine the exact model of that trend or the exact structure of that trend. Or, the presence of some stochastic trend. So, a stochastic trend means a trend that depends on time.

So, such a trend is called a stochastic trend. So, trends could be of two kinds. So, deterministic or stochastic. So, these three points kind of amount to non-stationarity in the time series. Okay, so now the last section of this lecture would be discussing some tests for stationarity.

So, are there some hypothesis tests, or can we perform some hypothesis testing in a scenario or situation to test if a time series is indeed stationary or not? So, in literature, different people have come up with their own tests. So, for example, the first one is called the augmented Dickey-Fuller or, in short, ADF. So, the augmented Dickey-Fuller test or, in short, the ADF test. So, typically, what this test does is that it tests if a unit root is present in a time series or not.

So, this ADF test tests if a unit root is present in a time series or not. And a short while back, we have seen that if there is a unit root, it implies non-stationarity. So, whenever you have a unit root in the polynomial equation, comprising the backshift operators and so on, it implies non-stationarity. So, what exactly is the hypothesis here? So, the null hypothesis is that the series is not stationary, whereas the alternative is that the series is stationary.

So, for some reason, if you reject the null, what does that mean? So, if you apply this ADF test and you reject the null hypothesis, it means that the series is, in fact, stationary. So, if you reject H0 or if you reject the null hypothesis, it means that the series is, in fact, stationary. So, all these tests, by the way, can be performed in any software, including R. So, you have a function called ADF.test. So, during one of the practical sessions, we will take it up, but just to reference it here.

So if you apply this ADF test on any time series. So let us say your data could be temperature data. Just an example. So let us say the ADF test is applied on temperature, and then it will give you a p-value. So any ADF test, if you are using software, produces a p-value.

And then you can compare that p-value with the significance level. So if the p-value is less than alpha, then you reject the null. Otherwise, you fail to reject the null. Make sense? So the first one is called the ADF test.

Then the next one, the name is slightly difficult. So even I won't pronounce that. So in short, it is called the KPSS test. So KPSS are short forms for or initials of four people. And then they produced a test as a combined kind of research paper.

And then, what exactly is this test? So, this test, the null hypothesis is that the time series is stationary around a deterministic trend or mean, against the alternative that it is non-stationary. So, the idea of the ADF test was to check for unit roots, while the idea here is to check for a deterministic trend component. And then, the null hypothesis is that the series is stationary. The alternative is that the series is non-stationary.

So, this is the KPSS test. Then, the next one is the Phillips-Perron test, or PP test. So, this is kind of similar to the ADF test, but it makes different assumptions about the error terms and is more robust in the presence of autocorrelation and heteroscedasticity. So, heteroscedasticity means nothing but changing variance. This is a technical time series terminology for describing changing variance.

So, changing variance means heteroscedastic. And then, if the variance is not changing, if the variance is constant, we call that homoscedastic. So, again, similarly here, the null hypothesis is that the series is not stationary, and the alternative is that the series is stationary. So, again, this PP test is kind of similar to the ADF test, but has some different assumptions about the error terms and, generally speaking, is more robust in the presence of autocorrelation and if the variance is changing or heteroscedastic in nature. And then, the last test is called the variance ratio test.

So, what exactly do you mean by the variance ratio test? This kind of test for a random walk hypothesis indicates non-stationarity because, as we have seen in previous lectures, any random walk is not stationary, right? So, it kind of compares the given series with a random walk. So, if you see a resemblance with a random walk, then one can actually

argue that the series is not stationary, right? And then, this variance ratio test kind of produces a ratio.

So, if that ratio is significantly different from 1, this in fact suggests that the series is not a random walk. And if the series is not a random walk, it means that the series has to be stationary, right? So, the null hypothesis here is that the series is not stationary. And the alternative hypothesis is that the series is stationary. So, essentially, through all these different tests—ADF test, PP test, variance ratio test, KPSS test—one can actually perform a hypothesis test, which is a much more formal test.

This is a more formal approach than simply commenting based on some graphs and so on, right? So, if one is able to perform any of the formal tests—ADF, PP, etc.—then one can actually gain much more confidence in whether a series is stationary or not. Alright, so now, probably in the next session, what we will do is converge to the SARIMA model that we talked about earlier. So, we will kind of define the SARIMA model. We will again link it back to why exactly you need the extension from ARIMA to SARIMA, right?

And then, how do you expand that SARIMA model? What are the orders of a SARIMA model? Along with some examples. Thank you.