Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 01

Lecture 01: Time Series Introduction

Hello all, welcome to this first lecture in this course on time series forecasting with applications in R. So, in today's first lecture we will focus more on introducing what exactly is time series and how time series data is kind of different from any other data that you might have seen in let's say any of the other statistical courses. For example, let's say inference or be it regression or be it hypothesis testing and so on and so forth. And this first session we will try to cover a lot of examples of different time series coming from different areas like ecology or be it finance or be it climate studies and so on.

So here if you see we have a stock price of Apple and this is a very classic example of time series and all of you can appreciate exactly as to how time series behaves. So, I think this is slightly old it dates back to '21. So, August of 2021 to the current month and here you can see the behavior of the Apple stock as time moves and down here you can also see the volume. So, stock price, when somebody talks about time series, they kind of appreciate time series with certain stock price. And of course, like I said, down the line, we'll talk about multiple other examples also. So, before we give out any examples and time series, I'll give you a small question to all of you.

So, in front of you, you can see two options. The first one is a very simple one. So, let's say you have X_1, X_2, ..., X_{100}. And these are, let's say, some random variables. And assume that these are the outcomes of some dice roll in, let's say, 100 rounds.

And as opposed to that, if you focus on this second example here, let's say X_1, X_2, ..., X_{100} be the closing prices of a particular stock again for, let's say, some 100 consecutive days. So now the important question is what exactly is the difference between these two options? So maybe I'll give you half a minute to think. Okay, so I think if you consider the first option, again, all of these X_i are random variables. So, if you roll a die, then one can expect that any number between one to six can occur, right?

So, all these occurrences are completely random in a sense that they're kind of independent to each other. Right? But again, if you focus on the second series here, which is let's say X_1, X_2, ..., X_{100}, where all of these random variables are some closing prices of a stock, then one can actually assume that there has to be some inbuilt pattern in that. So, for example, X_2 and X_3, or let's say X_3 and X_4, they need not be completely independent in a sense that all of these random variables, they kind of depend on the timeframe or the time factor there. Okay. So, this is I think very important and the foremost distinction between a time series data set and any other data set for that matter. So, before you proceed, I'll give you a very quick example.

So, let's say in almost all the other courses, you must have seen random variables which are written down as X_1, X_2, X_n, okay? And further, what we assume is that they are coming from, let's say, some distribution. So, I'll take normal, okay, with some mean mu and some variance sigma square, okay? And the foremost assumption that we make is all these are IID. So IID stands for independent and identically distributed.

So, I will say that this is a classic example of a data set. So, I'll write down classic data set. But then as opposed to that, if you transition into a time series, then how exactly a time series data is different is due to this time frame here. So, I will mention X_t and then the next observation would be something like X_{t+1} and so on up to whatever time frame you want to consider.

So, let us say the last one could be something like X_{t+10}. Okay? So, if you see here closely or if you observe closely, the first set is completely independent of each other. So, any pair of random variables are independent, but when you talk about a time series set or time series observations, for example, X_t, X_{t+1} up to X_{t+10}, then all these individual random variables need not be independent in a sense that all of these are dependent on that time frame t for that matter, ok. So, of course, in the later slides, we will see that how exactly you can change this time frame 't'. So, we can have different situations pertaining to different focus areas. For example, 't' could be daily or 't' could be weekly or let us say 't' could be monthly or 't' could be annual, right?

So, depending on what kind of data structure you have, one has the liberty to change that time frame notation 't'. So just to sort of summarize exactly what we discussed in the previous slide. So, what exactly is time series? And the second point is how time series data is different from any other data? So, the first point is classical inference.

So, classical inference covers any other data set which you might have used in some other course, let's say regression for example, or statistical inference for example, right? And as we discussed earlier, there the data is sort of assumed to be completely independent and identical. So identical means that all the random variables have to come from some fixed distribution, let's say binomial or exponential or normal, et cetera. But the moment you talk about time series dataset or time series observations, the first point is that any time series dataset is ordered. So, there is a certain ordering in all the observations, right?

So, let's say X_1, then X_2. So X_2, when you talk about X_2, you kind of assume that X_2 succeeds X_1. So, let us say a simple example could be temperature data. So, let us say X_1 could be today's temperature, X_2 could be tomorrow's temperature, X_3 could be day after tomorrow's temperature and so on. Alright.

So, time series observations, there is a certain chronology in that. So, data set is ordered according to time or we can say that the observations are chronological. And just to summarize this question in red, it says that, hence, "is time series data independent?" So again, the clear answer is no, right? So, time series data is not independent because all the individual random variables, let's say X_t, X_{t+1}, X_{t+2}, etcetera, they sort of depend on that time frame 't'.

Okay, so before we proceed, I'll give you a short overview of what kind of data types we have. Alright. So, the first very important data type is called as cross-sectional data. So, I think cross-sectional data is any kind of data set that you might collect in a usual day-to-day situation. So, let's say, heights of students or weights of different students or ages of different students. But the key definition here is observations come from different individuals, but at a single point of time.

Okay. So again, just to repeat, any cross-sectional data, all the observations have to come from different individuals, but at a single point of time. So, the single point of time is sort of important here. So, at any given time point, if you're collecting data set or if you're collecting observations from different individuals, we'll say that the data set is cross-sectional. Now of course, in the next slide, I'll give you some examples. But then the other kind of data set is time series data.

And then focus of this entire course, per se, is especially on the second point, which is time series data. So, what exactly is time series data now? So as discussed earlier also, that time series data is nothing but set of observations collected for a particular individual. So, you are kind of fixing one individual or an entity over different time

points. Alright. So, the differentiating point between cross-sectional and time series is cross-sectional is collected for different individuals at a single point of time, whereas time series data is collected for a particular individual or an entity over different time points.

So now, I think we'll discuss some examples of both kind of data sets. So, the first one is cross-sectional. And then here you can see that you have examples coming from different different areas. So, the first one is let's say maximum humidity levels at 20 different locations in India on a particular day. So again, just to highlight, so we're talking about maximum humidity at 20 different locations on a particular day.

So, the time point is fixed here. And we are collecting data observations from 20 different locations. Or the second example could be something like the closing stock price of 20 stocks again on a particular day. Or one can talk about, let's say, heights of students in a class measured on a particular day. All right, so again, just to repeat that any cross-sectional data, the key element is to fix the time point and then collect the observations coming from different individuals or different locations or different entities, et cetera.

Now, as opposed to that, the second kind of data set is called a time series data set. So, the first example could be a very simple one, something like daily maximum humidity levels for a month. So again, just to elaborate on the first example here, so here we are not fixing the time point, so we are kind of transitioning the time point over a month but we are talking about the daily maximum humidity levels maybe at a particular location. So here the entity or the individual or the location is fixed, whereas we have to change or keep on changing the time frame.

Or the second example could be something like the closing stock price of a stock over six months. Or the third one could be something like the quarterly student enrollment in a college over five years. So, again just to summarize very quickly. So, differentiating point between cross sectional against a time series data is that in cross sectional data, we sort of keep the time frame to be fixed whereas, we aim to collect observations or data sets from different individuals or at different locations or from different entities. Whereas when you talk about time series data, we are kind of shifting the timeframe and we are collecting observations over a wide range of a timeframe for a particular individual or a particular entity.

Yeah, so I think this slide kind of summarizes what we talked about a short while back, that time series data focuses on the same variable, over a period of time while cross-

sectional data focuses on several variables at a single point of time. So, I think I'll give you one short description just to elaborate this point more, because understanding time series and exactly understanding as to how it is different from, let's say, a cross-sectional data is kind of really important. So even for all the other slides that we'll talk about maybe later on, in all the other lectures, the focus would be entirely on time series dataset. So, if you're not very comfortable on understanding the difference between a time series data and a cross-sectional data, then I think there could be some problems.

So just to give you a very simple example as to how a time series data looks like. So maybe notation wise, we talked about a short while back. So, let's say X_t, X_{t+1}, then X_{t+2}…. So here again, if you notice, since this 't' is fixed, right and then since this random variable X is fixed, what we are doing essentially is that we are collecting observations from a single individual.

So let us say X, but we are kind of changing the time frame. So let us say 't', then 't + 1', then 't + 2', then probably 't + 3', etc. So let me mention that this is a time series data example. Whereas in a cross-sectional example, I will write down X_1, X_2, X_3 up to let us say X_n. So, all these random variables, let's say you assume that all these Xi are different individuals.

And we are talking about finding the heights of all these individuals. So, let's say something like H1, H2, H3 up to Hn. Okay. So here, what exactly is the series H? So, the series H could be heights of different individuals, but at a single time point.

So let me write down that these two pertain to not a time series data, but cross-sectional data. Okay. Now just to level up, so we have a data type which is called as panel data. or longitudinal data. So, what exactly is that?

And then how exactly is a panel data different from let's say cross sectional or a time series? We'll discuss it now. So, panel data, what it means is observations on different cross sections over a time. So, one can actually assume that a panel data is sort of a combination of cross-sectional data and a time series data. So let us say, think of a very simple example which is mentioned here, the first one.

So let us say the annual cancer mortality rates of different Indian states during 2015 to 2023. So, couple of points to mention here or couple of points to note here. Firstly, the observations come from different Indian states. So, we're not talking about a particular individual or a particular location. OK.

And on top of that, we're also changing the time frame. So, during what period? So, during 2015 to 2023. So, panel data is nothing but a combination of different cross sections over a time. So, let's say a simple example pertaining to the first example could be we can talk about different states, for example, Maharashtra or Delhi or Tamil Nadu or Jharkhand, etc.

So, what exactly are the annual cancer mortality rates in these locations over that time period? Or a second simple example to understand this panel data idea could be something like yearly sales of 10 companies over 10 years. So again, the same structure can be seen here that we are not talking about a particular individual here but sales of 10 companies over a wide time span which is 10 years. Or there could be one more example which is let us say daily maximum temperatures of 5 cities in India over a year. So, if you want to again link this panel data to either a cross sectional or a time series, again just to summarize very quickly is that panel data is nothing but a combination.

So, think of a situation where you have different cross sections spread over the entire time frame. So now I think we'll talk about what exactly are the steps when it comes to analyzing a time series data. So, I think since we are just starting and this is the very first lecture. So, I have highlighted a couple of very important points because there has to be some structure to analyzing a time series data up to the endpoint. So, I think the very first point is you have to plot the data.

So, if you gather any time series data set or you gather any time series observations, the very first step is to plot the data. And analyze it to extract meaningful information. So, I will give you a very simple example. Let us say if you want to talk about Google stock price. So, the X axis and the Y axis.

So, the X axis is naturally time. And the Y axis could be let us say Google price. Google stock price. And again, this is a very hypothetical example. So, let's say the price moves something like that. Okay?

So, since we've plotted the data, we can actually see that what patterns they're following. So, what patterns the Google stock price is following. So just to elaborate more on this point, so one can actually see that, So, initially in the very first stage, the price is increasing. So, we have a sort of a trend here, right?

And then probably in this period, the price is not moving much. So, price is kind of stagnating. And then again, the price sort of jumps here. So, by plotting the data, one can

actually analyze the data much more easily and then try to find out the underlying patterns in the data set. So, this should be the very first step.

The second idea is to study the past behavior. So, what exactly do you mean by this? So, study the past behavior means that how has the Google price behaved in the past? Or how has the Google price behaved in the history? Because that will give you some idea as to how it will perform or might perform in the future.

Now the third point is to identify the underlying patterns and trends. So, trend is a very useful terminology when it comes to any time series lecture or time series course. So, the idea of trend is very simple. So, trend is any upward or downward movement of an entity. So, for example here, again coming back to this Google stock price example.

So initially we saw that the Google stock price was uptrending. So, you have a trend here and then you don't see any major trend here but again down the line you again see a trend here. So, by this we see that the Google stock price was kind of trending initially but not trending in the middle and again trending towards the end. So, one has to actually identify all these patterns.

So, do you have a trend in the data or do you have any repetitions in the data set? So, all these are kind of patterns which are very useful to analyze any time series data. And I think the last point is use past data for forecasting. So ultimately, the goal of any time series course or time series analysis is forecasting. So forecasting is exactly similar to prediction.

So, I'm not sure if you've done any course on regression, but in regression, we have a similar terminology called as prediction. So, predicting something in the future. Similarly, when we talk about any time series lecture, we will say that we have to forecast something in the future. So, for example, forecast in the Google price for the next 10 days or forecast in the Google price for the next month. Alright. So, using past data for forecasting is the ultimate goal.

So broadly speaking, these are all the steps involved when it comes to analyzing any time series data. So now that we have discussed some examples of time series data set, for example, the Google stock price or the Apple stock price is there on the very first slide. And after understanding the fact that how can you differentiate between a time series set of observations vis-a-vis a cross-sectional data or any other data set that you might encounter elsewhere. So now we'll focus majorly on a few other application areas of time

series. So, I think this slide is kind of important to kind of gauge on the fact that in what all areas can time series be applied or in what all areas can the application of time series can be seen predominantly.

So, for example, the first one is a very simple example to start with. So, let's say retail stores. So, retail stores can use it to forecast sales. So, let's say D-Mart or Big Bazaar, right? So, all these retail stores, might use the time series analysis to forecast their sales. Now, when you talk about sales, the time frame here could be slightly different according to what goal the retail store has. So, for example, forecasting daily sales or forecasting monthly sales or forecasting quarterly sales or forecasting the annual sales, right? So, depending on what the goal is, the time frame can be changed here. Now the second example is let's say energy companies.

So, energy companies use it to forecast reserves or their production or demand or supply or prices etc. Educational institutions might use time series analysis to forecast student enrollment. So, let's say an institution like IIT Bombay can actually apply some time series modeling or time series forecasting techniques to kind of forecast as to how many students would enroll in the next semester or next year, etc. Or for that matter, international financial organizations, let's say such as World Bank or International Monetary Fund, might use time series analysis to forecast inflation or any other economic activity.

The next example is let's say transportation companies. So, transportation companies might use time series analysis to forecast their future travel. So, the problem statement could be something like how many people would want to travel let's say in the next month at a particular destination or how many people would want to take a vacation in the coming summers or in the coming winters. So, it's a very valid problem statement. Or let's say banks and lending institutions might use it to forecast new homes purchases

Or let's say venture capital firms might want to apply some time series techniques to forecast their market potential and vis-a-vis evaluate the business plans, etc. Or the last one is again a simple example. So, let's say weatherman might use time series analysis to predict precipitation, temperatures, humidity levels, etc. So, I'll say the time series finds its application in numerous areas such as let's say finance or climate studies, environmental sciences, etc. So now I think in the next couple of minutes or so, we'll discuss what exactly do you expect in this course.

So, I've just prepared two slides on this and we'll go week by week just to understand the fact that what advancements can we talk about here. So naturally in the first couple of weeks, we have a small introduction, which we are kind of continuing now also. And then there's a concept called a stationarity in time series. So, we'll discuss that. Then we'll talk about some basic time series models to start with.

Moving average, autoregressive, etc. And then we'll wrap up the first couple of weeks by talking majorly about non-stationarity. So almost 90% of the examples that you see in the real life are not stationary. So then when you transition into the other weeks, so weeks three and four, so we'll talk about some tests for stationarity, some other advanced models such as ARIMA, SARIMA, and then model identification. So how to identify the correct model?

And then later on in weeks five and six, we can focus more on, let's say, forecasting methods than comparing forecasts. So how would you compare two different forecasts which you've applied for the same time series and then pick the better one? Then you have a topic called fractionally integrated processes. So, we'll spend some time there. And in week seven and eight, we'll focus more on multivariate processes.

So multivariate time series, then co-integration. So how are two time series co-integrated? And some causality analysis. Then down the line in weeks nine to 10, we'll focus on four-year transformation. So, I think this week would be kind of interesting where we'll transition

into a slightly different angle of looking at time series data sets, which is Fourier transformation or spectral density and then volatility modelling. So, models like ARCH or GARCH, they'll appear here. And in the final weeks, we'll spend some time on non-linear models or let's say Markov switching models and then ultimately, we'll talk briefly about machine learning models. So, these days, people are more interested in kind of mixing or applying some machine learning modeling techniques into, let's say, analyzing some data sets and so on and so forth.

So, we'll spend some time on machine learning models and applications using a time series data. So, let's say neural networks or decision trees or reinforcement learning, etc. So, I think this is a very brief outline about how the course will progress down the line And of course, in each of the weeks, there'll be a small component of some practical things in R using the R software. And again, in the subsequent lectures, I'll give you a short overview of R as well.

So, if somebody who is not very confident in applying R or analyzing using R, they should not worry. So, we'll discuss that in brief. Thank you.