Transcriber's Name: Prabhavathi
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology-Kharagpur

Lecture-44
Auto-Regression Analysis

Hello guys. So, today in this module on relation analysis, on the module on relation analysis, we have basically learned correlation analysis, we have learned relation analysis, I means regression basically, we have learned correlation analysis, we have learned regression analysis. Today we will be learning a special type of regression analysis that is called auto regression analysis. (refer time: 00:50)

Now the question is why it is called special. Why I told that it is a special? It is not all special, but why I told it is a special type of regression analysis. That is basically this is here a regression analysis will be doing on a different type of data. What is the different type of data will be doing this regression analysis and the time-series data okay. So, now what is time-series data I will come to it shortly.

So, and now when we do correlation on this time-series data we do not call it the simple correlation analysis, we call it autocorrelation. Similarly when we do regression analysis on time-series data we just do not call it regression analysis, we call it auto regression. So, basically in this lecture, we will be learning what is time-series data; we will see autocorrelation, auto regression and will also see some of the applications of this time-series data okay. So, now first what is time-series data? (refer time: 01:45)

First let us understand what is time-series data. Basically okay I will go to the figure. First if we consider time-series data it is like when we tell data what is the data? It is data means we collect we collect data on some elements or some on some variables like rainfall. So, it is a data we this is data of some particular variables right. So, like our date of birth and so just just if I write 19 say 1990 what does?.

It does not mean any meaning or just if I write 22nd July, it does not carry any meaning. What this data is of what? So, we collect data on some or some element or some ahh observation or some variable. So, when we collect data of some observations repeatedly we do repeated measurements, over a period of time when we collect data of some observations, over a period

of time we do repeated measurement then we call it a time-series data.

Like this repeated this repeated observation can be anything. Like all of you I think you know when we when I told you the time-series data like rainfall direction, where time-series data. So, yeah there are simple very time. some simple time-series data may be like ahh if I want to find out the sale of a particular product in a shop, sale in hourly basis. If I can calculate the sale of a particular product in hourly basis or daily basis, that also becomes a time-series data.

When I am Basically when I am collecting data of some observations through repeated measurement of time, then I call it a time-series data okay. So, now basically so and okay Now you got an idea what this time-series? This is an example of a time-series data what I have given here in this picture. See it is a rate of price inflation. Price inflation is also time-series data; because we are collecting the data of the price the inflation of the price repeatedly we are collecting this data.

Repeatedly we are calculating this, taking repeatedly we are measuring this data on a over a period of time or in a particular time period. So, here suppose we have calculated the price inflation, we have calculated the price information for a couple of years from 1960 to 2005. And 1960 means we did not just take the data of 1960, if you can zoom it further then you will see we have this inflation data for each day of 1960, similarly each day of 1965.

We have calculated this. We have measured this inflation data for each day of 1970, each day of 60, 61, they base in 60, 61, 62 for each day. So, does if you zoom it further you may see it this way, but of course in this slide I could not keep it that way. So, this sort of data we call it a when we are repeatedly measuring this observation over a time period, we call such type of data we call it a time-series data.

So, here in the the vertical axis, what is that here we are taking the percentage. So, it can be in a different format. Now here we are taking the percentage. What is the rate of price inflation for each year? So, it is here we are giving in percentage. This percentage may be it may be negative inflation, it may be positive inflation and we get in percentage okay. (refer time: 04:58)

So, some other examples of time-series data like okay aggregate consumption and GDP for a country, for example 20 years of quarterly observatories. Each year we take 4, 4 observatories, 4 observations means after we take observation for each 3 months. So, how many total observations we will get? For 20 years, we will get 80 observations. This is also in time-series data.

What to say We are collecting data of this aggregate consumption and GTP and we are doing sort of repeated measurements for this, repeated measurements so, how much how much repetition is that each year we are calculating it 4 times, after each 3 months we are taking the value. So, this is also a time-series data. Then again Yen per Pound, Yen per Dollar, Pound per Dollar, Euro per Dollar exchange rates.

All this exchange rate if we take daily data for 1 year this is also a time-series data. This may be if you just take it for one it may be very small time-series data. Usually time-series data we consider here for many years. Then cigarette consumption per capita in a state, by year for many number of years. This is also time-series data. Rainfall data, rainfall data is a very good example of over time-series data.

Rainfall data over a year or a period of years, this is an example says when time-series data on does not need to be a very great thing it can be very small thing also very small concepts like like sales of tea. Sales of tea from a tea shop in a season that can also be also time-series data. If you consider the sale of tea each hour or sorry let us not take each hour if we consider a sales of tea each day in a particular season. That can be also a time-series data okay.

Not okay These are the examples of time-series data. Can you give me an example, which is a which is which we do this repeated measurement, but we which I will not take repeated I should not say it is a repeated measurement, but ahh data which varies over time. But then we do not call it as a time-series data, can you just give such example? okay ah okay Once this example is video data, audio data these are not time-series data okay. So, why they are not times this data? okay I will come to that later OK. (refer time: 07:16)

Which are the following graph is due to time following graph is due to the time-series data here. They are given 4 graphs as this graphs they represent their time-series data, well unfortunately no. Non or the graph represent the time-series ticket data, because here we see is 200, 400, 600, it does not seem to be like a periodic data. Means we are taking the where it does not specify that it is we are taking measurement repeated measurement over a period of time, none of the data. Here also is 1, 4, 7, 10 same. So, none of this figure is that it is represents a time-series data okay. (refer time: 07:57)

okay So, now this time-series data it has many effects it has sorry it has many uses. It is a very, very useful data I am like we can use it in many different things. First one of the very good uses of time-series data are which almost all of us know is that as forecast model. How can we focus the ah whether it will rain, how much it will rain next year. Before the monsoon comes people forecast that it is not general people that people who deals with this they really focus, they say

how will be the monsoon this time.

So, ahh so this this type of forecasting model what will be the rate of inflation in the next year. This All this this forecasting model it is the one some it is a really very good application of time-series data. So, basically how it is forecasting means we are basically we are predicting. So, predicting remains us we are doing regression linear regression. So, this is when we talk of focusing it is nothing but linear regression only.

We are The thing given an independent variable, based on the independent variable we are giving the value of a dependent variable. Here the the what is the dependent variable, dependent variable? is a In this stage dependent variable is a inflation in the next year. And what is the independent variable? Independent variable the inflation rate of many years down the line maybe.

So, this is one of the uses of time-series data that is the forecast model and to estimate dynamic causal effects, dynamic causal effects very good. So, like one good example is like this this mobile service provider. There are different mobile service provider, so now the this this service provided, they introduce different schemes, they introduce different what to say pro schemes and they sometimes reduce the rate, they sometimes increase the rate and because of this what how how it affects the customer?

Basically what is the customer chance, because of this different proposal a different scheme they have introduced? So, customer chance means I think you know like what is the percentage of customer living that mobile service provider. So, because by analyzing those it can find out the causal effect means, because of this ahh thing it is having this effect. Because of introducing this scheme people are leaving the service provider.

It is finding out the causal effect. So, this is time-series data also used to estimate the dynamic causal effect. Like again here given a good example, if the rate of interest increases what will be the effect on the rate of inflation and unemployment in 3 months, in 12 months? If the rate of interest increases, if the rate of interest increases then what happens people will stop expanding their business, people will stop opening new business.

And then what happens, if people become very conservative. Then what happens unemployment will increase unemployment will increase, inflation will increase is not it? So, basically how will you know? We will be able to predict by seeing the data of previous years. How this is happening having an effect on this. So, these are causal effects. But OK this is ahh This is from the perspective of a business perspective.

Now even we can also see it is a relation it is application and reliability analysis also. Reliability analysis, suppose given a system given a system and we have the data of system unavailability on a daily basis or on a monthly basis or weekly basis. System unavailability means when a system goes down it is obvious each and every system is at some point or other will go down is not it?

And suppose system is in continuous use and the system goes down, there has to be some repair action. When the repair action is done, then again the system will be up. So, when the system goes down till it is corrected, till it is repaired it is the system remains unavailable. So, this is the unable unavailability time. So, if we have this information what is the on an average what is the unable unavailability of in a day or in a week whatever it is.

And suppose I need to use this system for a very critical application, where my unavailability should be a particular value, it cannot exceed this particular value. Then I will be able to analyze this my past data and I will be able to tell whether this is suitable for that application or not. Because if the unavailability time increases, then definitely I will have to work something and maybe I will have to put more people on the repair team.

Or maybe I will have to use some advanced repairing actions. So, that my unavailability get reduced so there okay. Those are different management; my thing is that first if I want to find out ahh based on the unavailability data of past history I will be able to predict how much it will be system will be unavailable for how much period of time or maybe this is from the unavailability perspective.

Or maybe we want to similarly from the failure perspective also. If I know the daily basis the system fails how many times. So, that if I take the data for a year then basically I will be able to predict for the next year okay. So, these are some of the dynamic causal effects then time dependent analysis. Time depend analysis like rate of inflation, unemployment it can be observed over a time period. These are the different uses of the time-series data okay. (refer time: 13:28)

So, we have seen the uses of different uses of time-series data. Now how we will get this it is this use this application to get this application of time series, about data we need to we need some sort of modeling right. We will have to model this data then will be then only will be able to get those uses. So, first thing when I tell it is a forecasting model. Then definitely it is a regression model, will have to have a regression model, then only forecasting is nothing but prediction right.

So, if it is if we need if you want to use this time-series data as a forecasting model then we will have to basically find a regression model. Similarly if you want to find out the correlation over time. Correlation over time that means we will have to find an autocorrelation among this data. Autocorrelation it is also called serial correlation. We will have to find out the correlation among the, suppose I am I want to find out the correlation of this year data with the previous year data.

I want to find out the correlation of this year data and previous to previous years data. Basically I want to find a correlation between say 2020 data and 2022 data. I want to find a correlation between 2022 data and 2010 data. So, this way if I find a correlation, this is also called serial correlation. So, I have to find out, I will have to need the correlation model also. So, then again you can model to find out the dynamic causal effects also.

So, this for this different application basically we need to model this data. So, in this lecture we will not be discussing about the dynamic casual model, for dynamic causal effects, will be mainly our focus will be on the forecasting model. (refer time: 15:04) So, now ahh how to model it, like here since suppose there is a time-series data what it is given? Dollars per English pound, so this data ahh this time-series data it significance it signifies dollars per English pound.

So, if this is the value given to us, if this is the data this is the data basically. And this is means 2007 means not one value, basically 2007 means it is maybe data is collected on each day of 2007. So, that is why we can see it in a continuous form, because if we see the data as data is actually discrete. Here we are seeing it in a continuous form. Why because we are collecting the data for each day.

So, when we have this sort of data then can we predict the trend at a time say 2017. In 2017, what will be the value of the dollars per English pound can we predict the trend? Yes, of course we can predict this trend. We will be using regression, auto regression. We would not tell it is regression, we will tell as auto regression okay.

So, now to we need to understand what is auto regression? So, to understand auto regression basically, we will need to do lots of calculation. And this, what is calculation? We will need to understand autocorrelation. Basically, as I told you our focus here is auto auto regression, but for doing auto regression we need also to know auto correlation. So, we will have we will learn that also, but before that let us see some concepts and notation before we go to those modeling in details. (refer time: 16:33)

So, what are the some notation, when we tell Y t is the value of Y in a period of t. Say Y 2022 rainfall data, that is rainfall data. I know that is that if I use the variable Y that Y 2022. Now this Y t means they can be again is data set for each when that means there can be data for 365 days okay. So, data set this is the T observation on the time-series random variable Y. For 1 year, Y 1 to Y 365, again for the next year Y 1 to Y 365.

So, this is one notation. Then I see the example here Y t denotes the daily, weekly, monthly rainfall in year. More precisely, for the year 2021 say 365 days data. So, the rainfall data from 2011 to that is, for the last 11 years data okay. (refer time: 17:30) Then some assumption for time series, then we can call that the data is a time-series data, where it has to satisfy certain assumptions and what is that.

We consider only conjugative, evenly spaced observation. So, we consider for any forecasting model, for any correlation model our time-series data needs to be conjugated evenly spaced observation. For example, say monthly data in 2010 to 21 for each year and without any missing month. If you are considering monthly data from 2010 to 2021, then we should have data for each month from 2010 to 2021 each month of 2010, each month of 2011, each month of 2012 up to each month of 2021.

There should not be any missing month and there should not be any other data. For example, on daily basis for a year is admissible. We are taking monthly data. So, we do not need another there should not be any other data like daily data daily data. We It is not admissible. Then, another one property with this time-series data should follow; the time-series it is is stationary if it is probability distribution does not change over time.

Remember this audio data; video data I told it is not time-series data, because of this property. This time-series data audio data this is does not satisfy this stationary property. But that means what its probability distribution does not change over time. Then now means like say the rainfall data in 2020 we found it follows a normal distribution, in 2021 is found is follows an exponential distribution.

No, data may vary but then there is probability distribution it does not change over time okay. So, then we call it just stationary. A time-series Y t is stationary if its probability distribution does not change over time, that is, the joint distribution of this does not depend on i okay. It does not In rainfall data it does not depend on each year. In 1 year, we will see that rainfall follows a normal distribution, other exponent distribution other year some other distribution.

No, it is it follows a particular distribution then only we can consider the time-series data. Then

only we can use it to use it for for modeling those data. If the problem if the data is not stationary, we cannot model it we will not be able to predict it properly. So, stationary property implies that history is relevant. In other words, stationary requires the future to be like past in a probabilistic sense of course okay.

Auto regression analysis assumes that Y t as both uniform and stationary. So, if you are interested in doing the regression analysis, we are assuming that Y t is both uniform and Y t is stationary. If With this assumption only we can carry out the auto regression analysis. Else if, it is not stationary and if it is not uniform, we will not be able to carry out auto regression analysis, the result will not be correct okay. (refer time: 20:29)

So, now there are 4 ways to have the time-series data for auto regression analysis. For auto regression analysis there are these are the different 4 ways. How our data will be. One is we call it lag. The first leg of Y t is Y t 1. The rainfall data of 2022, it is first lag is rainfall data of 2021. It is second lag, rainfall data of 2022, it is second lag is rainfall data of 2020, that is, we call it lag. The first leg of Y t is Y t 1, it is jth lag is Y t - j okay.

And tth lag is 2022 - 5 that is the jth lag. Then we have a term called difference. The first difference of a series, Y t is a change between period t and t 1. This is how we call it a difference. Log difference sometimes the data changes abruptly. So, it is very difficult to just to take the data. So, in those cases we use the log of it. Then we call it a log difference okay. Then again similarly we can have also percentage. Like in the inflation rate, we have used percentage. So, this is the percentage okay.

So, now we will see autocorrelations. (refer time: 21:48) The correlation of a series with its own lagged value is called autocorrelation, also called series correlation. Remember we have to use the term series correlations, earlier when I used when I told ahh this modeling of time-series data, so this is a series correlation. When we are doing correlation of a series with its own lag value, 2021 rainfall data I am doing correlation with two its own lag value it is for 1 year lag that means, with 2022 I am I want to predict the rainfall of 2023 maybe.

And I am sorry we are not predicting this correlation. I want to find out the correlation between 2022 rainfall data with 2021 rainfall data or I want to find out the what to say ah correlation between 2022 data with 2020 data okay. So that is called as the serial serial correlation. Here we use that and this lag value, this lag. So, formula for the jth auto correlations, this is the formula for the jth autocorrelation okay.

So, we found a covariance between Y t and Y of t j. So, remember how we found out the

correlation? Correlation is covariance of x i and y i divided by standard deviation of x into standard division of y is not it? So, same thing it is just what is x i here? X i is the time-series data Y t and ahh what is y i? Y i is this lag data. So, 2021 and 2020, this is 2021, it is 2020. It is just I am taking the example in the from the rainfall perspective.

So, it may be not in in terms of year, it may be in terms of months, whatever it is. So, this is how we found a jth auto correlation Okay. So, covariance of the Y t and Y t j is the jth auto covariance. We do not call it covariance and simple data we call it covariance. Here we call it auto covariance. (refer time: 23:57) So, how to find out the covariance between variable Y t and Y t 1, this is the formula.

Why it is x i = summation of i = 1 to n, because for that particular year or whatever it is there will be many data right. If you consider here maybe if we have day wise data so i i = 1 to 365 right. So, this is the this is how we find it auto covariance. (refer time: 24:26) So, now suppose the same example what we have seen the value of dollars per English pound. So, for is given that is suppose we have given a ρ 1 that means, between 2 consecutive years ρ 1 that is between two two consecutive 12 and 13, ah 14 and 15, 21 and 22.

So, for given data, ρ 1 is 0.84 between two given consecutive years. This implies that the we got this is the auto covariance value factors is ρ 1 is 0.84. This implies that a dollars per Pound is highly severely correlated. 0.84 is very near to 1, so this is highly correlated is not it? So, we call it highly serially correlated. If it is if its value is near to 0, then we will call it that is weekly serially correlated.

Similarly we can determine ρ 2, ρ 3 etcetera. P 2 means lag of 2 years, ρ 3 means lag of 3 years fine. So, now we will come to auto regression model. In auto regression model, basically why we have seen auto covariance, because we in auto regression model we will be to solve the to find in the regression model we will have to find the value of the parameters right.

So, once we find a value of the parameters then our job is done. We found a function. So, to find the value of the parameters, we will basically be using this auto covariance. This ρ, ρ 1, ρ 2, ρ 3 and all those stuff. (refer time: 25:53) So, auto-regression model and autoregressive model is used to model a future behaviour for a time-ordered data, using the data from past behaviours.

Concept of simple concept regression analysis we will be developing a model, which we call it ARM, AR model, autoregressive model. So, we will be developing AR model to model the future behaviour for a time order data using the data, from the past behaviours. Using the rainfall data

of past years will be able to predict the rainfall data of the next years. So, that is the autoregressive model.

Essentially it is a linear regression analysis, it is nothing but a linear regression analysis of a dependent variable using one or more variables in a given time-series data fine. So, one or more variables like if I want to find out predict the rainfall of 2022, I will be using 2021, 2018, 2019. If I use only 2021, this is 2022, if I use if I want to use 4 4 years, last 4 years 4 years down the line 2021, 2020, 19, 18 and 17. So, that is it. So, Y t is equal to function of this. How many years you want to use that is t - p okay. (refer time: 27:14)

So, a natural starting point for forecasting model is used to is to use past values of Y, that is, Y t 1, Y t - 2 to predict Y t. And auto-regression is a regression model in which Y t is regressed against it is own lagged values. So, we have already seen this. What is auto-regression model where, Y t is regressed against it is own lag values. What is the We call it independent variable is a regressor variable is not it.

So, Y t is here. Y t is the response. Y t is regressed against it is own lagged value. What are the regressors? Regressor is own lag value. The number of lags uses regressor is called the order of auto regression. To predict the rainfall of 2022, if I use 5 years before that, so that means, my order of auto-regression is 5. If I am using 2021 data 2021, 20, 2019, 18 and 17 then my order of auto-regression is 5 okay.

The number of lags used as regressor is called the order of auto-regression. In first order auto-regression denotes as AR 1, Y t is regressed against simple Y t - 1. In pth order auto regression denoted as AR p, Y t is regressed against Y t 1, Y t - 2 up to Y t p. fine So okay fine. We got regressive simple it is just a simple case of regression analysis. Regression analysis means we have to estimate the parameters. (refer time: 28:52)

So, in general pth order auto-regression model is defined as this way. This is how we define a pth order regression model. If it if we regret pth order means we will be regressing with the pth lag data. So, this is the intercept and then how many what to say parameter we have that depends on the order of the regression level. So, BHB is B 0 β 0, β 1 up to β p is called the auto-regression coefficient.

And epsilon t is the noise term or residue and in practice it is assumed to be Gaussian white noise. Epsilon t is just the noise term. It is assumed to be Gaussian white noise. What is Gaussian white noise? In simple if you see Gaussian white noise you do not have to go through such details. Just there is that this where its value is independent and its expected value of

epsilon t is 0, that is mean of epsilon t is 0. For your understanding that is sufficient. So, that is Gaussian white noise.

So, we have exp assumed that epsilon t is assumed to be Gaussian white noise. And this β 0, β 1 up to β p is the auto regression coefficient that means; we will have to find out the value of this coefficient. Now the task is to find out the best value. We have already seen in regression analysis our task is to find out the best value of this different parameters. How did you find out the best value for these different parameters?

Remember what we have done for linear multiple linear regressions how we have done? Simply we have to find out the sum of the squares error and then we have to differentiate first order differentiation. We get it to 0. And then we found out the value, that is what we have done it form. But here it is it will not be so easy, because the data is not so simple. (refer time: 30:39)

So, here actually basically just to find out the coefficient, one very efficient method has been very in general it is used, it is very popularly used it is called it is basically least square method. It is used under concept of really square method, same least square method what we have used in regression analysis. But least square method the way we do and where we find a expression how just we do it partial differentiation we find the expression to finding out the partial differentiation of age and finding the expression of this is a bit a bit complicated.

So, one very easy technique is Yule-Walker equation. It is called Yule-Walker equations. So, Yule-Walker equation basically it converts from the expression into ahh terms of this autocorrelation. This ρ1, ρ2, ρ3 is the, it converts this coefficient values, at it converts it shows a relationships with this coefficient values with this auto correlation values. So, Yule-Walker equation it is not possible to discuss at in this lecture, where it will be needing I will be needing a lots of board work.

And it will take from 2, 3 hours to explain the whole thing. I suggest you please Google it. You will be able to understand it and moreover if these things actually nowadays all software are available. Ready-made software are available and even you can write your own software using R so so, you just feed the data, you just select your technique and you will be able to solve it.

So, this is this Yule-Walker equation, just this is a this is the form where this we need to find out this values OK. P 1, ρ 2 are this is nothing but the correlations, ρ 1 is the correlation which just is next regress value. P 2 is this correlation from this value to 1 1 lakh okay. So, that way we we will having will have a system of equation, when we have a system of equation how to will be using we can use any numerical technique to ah what to say to ah solve this system of

equations.

Once we can solve the equation we will be finding out the values of β 1 to β p, if the order of auto-regression is p okay. So, now here I have used to solve for solving the equation many times I have suggested that you can use any numerical techniques. If By now if you have not learned a numerical technical (()) (33:00) some of the numerical technique methods which you really need it in many stage of your life will be needing I am sure there are some courses on that it is available and if it is not there it will be finding lots of material.

Please not only for this so in in-fact in many of the courses you will be needing how to solve a system of equations . Very not very simple equation, simple equation so you can easily solve it by if if not you can solve it by matrix method; you can solve it by Laplace method, but then if when you have to use numerical technique methods.

So, what are the different numerical techniques, how you can solve the equation. Please learn these techniques. I am sure some numerical technique methods courses are available. (refer time: 33:41) So, ρ i's are nothing but the ith correlation coefficient. So, from this equation we found out the value of β 1 to β p. And β 0 it can be chosen empirically, usually it is taken as 0.

Now we have this expression, we will solve this expression, we will find out the values of β 1, β 2 to β p. And once we found out the value we have the function. Once we have the function we will be able to predict given the what to say whatever order given 2 years data. Suppose that 2 years auto regression if given the 2 years data will be able to predict the third year data okay. (refer time: 34:15)

So, in this lecture we learned about the time-series data in what are the nature different natures of the time-series data and then we have also seen modeling with time-series data. We have seen auto-correlations and also we have seen auto-regression. In the next lecture, we will be learning logistic regression that is also an interesting topic okay. (refer time: 34:34) So, this is the reference. And thank you guys.