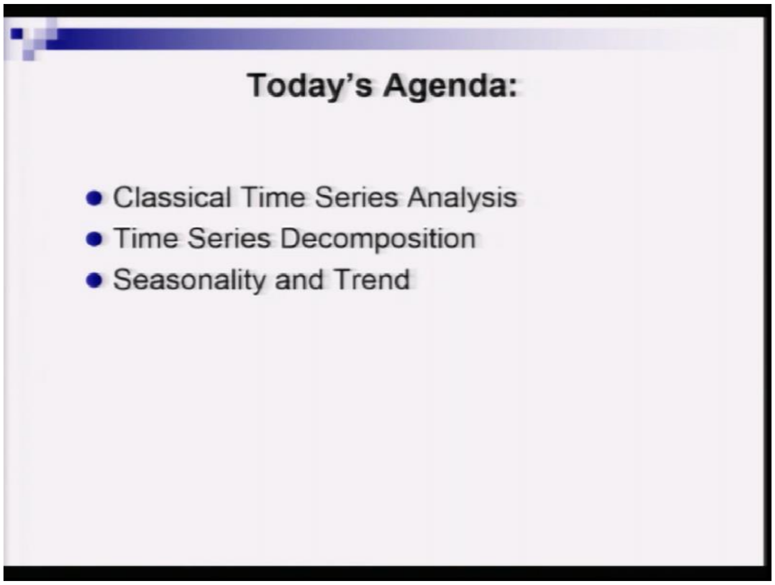
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-23
Classical Time Series Analysis (Part-I)

Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So today, we are going to start our discussion on time series data analysis. And let us have a look at today's agenda items.

(Refer Slide Time: 00:10)



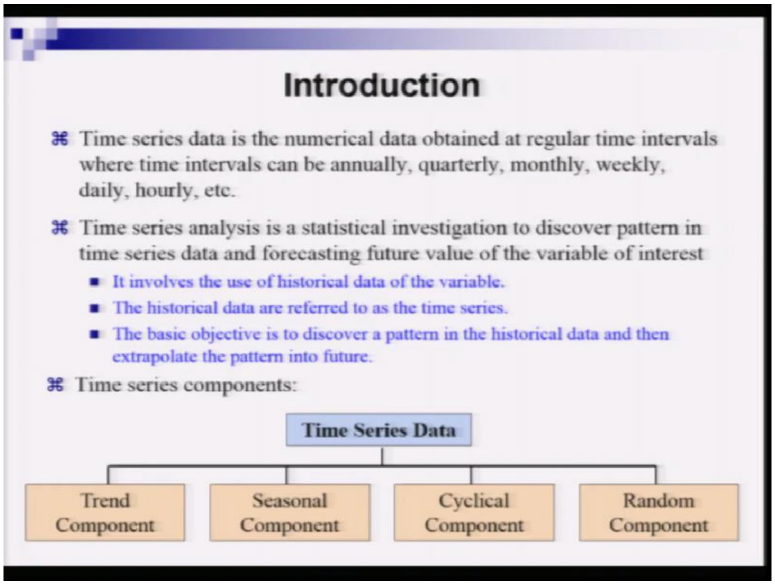
Today's Agenda:

- Classical Time Series Analysis
- Time Series Decomposition
- Seasonality and Trend

So, we are going to start with a brief description of classical time series analysis and here, in this class I am going to discuss the decomposition of a time series variable in two or three components. So, I am going to end today's lecture by discussing the case of seasonality and trend.

So far, we have dealt with the case of cross section data. Now, we know in this lecture and in the next lecture, we are going to discuss the case of time series data, where do you observe data over a period of time and before we go to the statistical modeling of these kind of data, let us have a brief idea about the time series data and, its components.

(Refer Slide Time: 01:15)



Introduction

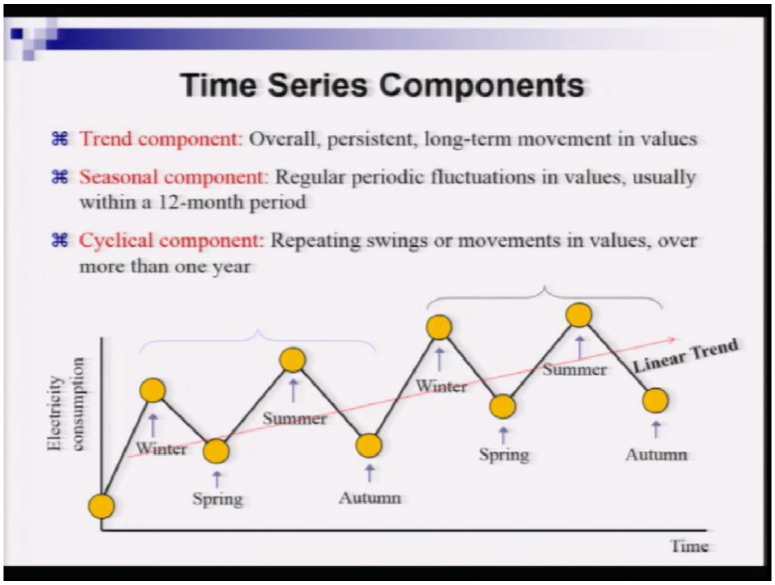
- ⌘ Time series data is the numerical data obtained at regular time intervals where time intervals can be annually, quarterly, monthly, weekly, daily, hourly, etc.
- ⌘ Time series analysis is a statistical investigation to discover pattern in time series data and forecasting future value of the variable of interest
  - It involves the use of historical data of the variable.
  - The historical data are referred to as the time series.
  - The basic objective is to discover a pattern in the historical data and then extrapolate the pattern into future.
- ⌘ Time series components:

So, a time series data is a numerical data set which is obtained at regular time intervals where time intervals can be annually quarterly, monthly, weekly, daily, hourly, in fact, in share market you can get per second or per minute data. So, what is time series analysis all about? So, it is a kind of statistical investigation to discover pattern in time series data; and finally, forecast future values of the variable of interest.

Now, the variable could be of any type, it could be economic variable like inflation or per capita income or it can be natural variable like maximum temperature in a day, it could be anything, but the most important thing is that this time series analysis involves historical data on the variable. So, you have to gather a large amount of historical data to conduct a time series analysis. And the basic objective is to discover the pattern in that historical data and then extrapolate the pattern into future values.

So broadly speaking, time series data has four components. And actually, the first three components are the deterministic components. And then, the last one or the fourth one is basically the stochastic component. So first, I am going to mention the deterministic components and they are of three types. One is called trended, the other one is called the seasonal component, and the third one is called cyclical component. And finally, fourth one, as I told you, it is going to be called the random or stochastic component. Now, let us have a look at these items or these components of time series data one by one.

(Refer Slide Time: 03:12)



Time Series Components

- ⌘ Trend component: Overall, persistent, long-term movement in values
- ⌘ Seasonal component: Regular periodic fluctuations in values, usually within a 12-month period
- ⌘ Cyclical component: Repeating swings or movements in values, over more than one year

So, let us start with the trend component of time series data. So, how can we define the trend component in the time series data? It is the overall persistent long run pattern that can be observed from the values. And what is the seasonal component? So, that is the regular periodic fluctuations in the variable values and this is to be observed within a 12 month period, within a year.

And then finally, we have the cyclical component, here we can observe repeated swings or movements in the values. But then we are talking about multiple years, so generally cycles are observed over a longer time periods say 18 to 20 years. But the length of the cycle is not equal. The idea of trend and seasonal component can be much more clear, if we look at the diagram that is there in the slide. So, here you concentrate on the diagram that you are seeing at the bottom part of the slide here.

I have taken a variable electricity consumption and I have observed data on this particular variable over a period of time, say I have data on 5 6 years and I have monthly data or maybe quarterly data. And that variable values I am plotting along the vertical axis and along the horizontal axis, I am measuring time, so basically, here the values could be no if it is monthly data, then we can have values of time like 1, 2, 3, 4, up to 12 then again 13, 14, 15, 16 then up to 24 then again 25, 26. Or if we have quarterly data then the values of time could be like Q1 can be represented by value one then Q4 can be presented by value four and then the first quarter of the second year that can be given the value five and correspondingly we can have different values.

So, for all these time values, we observe a specific variable value for the variable electricity consumption. And here in this diagram, I am showing you a very simple case. Suppose, we have data on two years and we have data on quarterly electricity consumption. So here, this quarterly electricity consumptions are plotted in these two dimensional plane and these are this orange color circles.

And here, you see that the electricity consumption is showing some long term trend. So, that long term trend is given by the red straight line that is passing through the scatter plot and I have indicated that as the linear trend with an arrow and you see, it is showing, an upward positively sloped trend line.

Now, if you look at the points in these this dimensional plane, you see there is a systematic fluctuation in the values within a year. So, let us start with the first-year case, so here you see, the first quarter is denoting the case of the autumn season. And then after that, you see there is a rise in the electricity consumption for the winter season and then it comes down at the spring season. Then again, it goes up for the summer season and then again it comes down to the autumn season.

And the similar pattern we observe for the year number two as well. So, there is a regular periodic fluctuation in the value for this variable, electricity consumption. So, that is the seasonal movement that this electricity consumption behavior is showing. So, why do we observe the regular fluctuations in an economic time series or in some time series variable? So, let us take the example of electricity consumption in a typical city in northern India, say Kanpur.

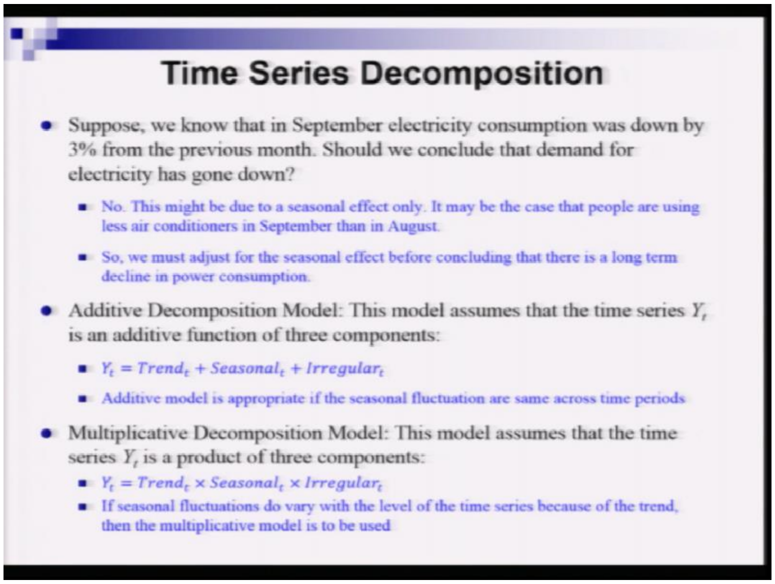
We all know that Kanpur has very extreme weather conditions. So, during the summer the temperature can be as high as 40 to 44 degree Celsius in a day and in the winter time the temperature can even drop down to 1 degree or 2 degrees Celsius, these are like very extreme temperatures.

So, of course, in the summertime, when temperature passes a critical temperature value say 35 degree or so, the use of air conditioning machines go up in the households and that is why the demand for electricity goes up heavily. And after monsoon comes with rain temperature goes down. So, of course the use of air conditioning machine is not that much required. So, there is a fall in demand for electricity consumption.

But again, in the winter times as the temperature goes down drastically below 10 degrees Celsius, then use of oil heaters go up in the city and then that will again raise the demand for electricity. So, you see that, due to weather condition actually there is a fluctuation or change in the demand for electricity consumption in the city. And hence there is regular fluctuating behavior that is observed in the households that will lead to the regular fluctuation or seasonal behavior of the electricity consumption variable over a period of time.

And finally, I would like to conclude by saying that as in any city population goes up with time. So, there will be an overall increase in the demand for electricity consumption, and that is basically measured by or represented by this upward sloping linear trend curve.

(Refer Slide Time: 09:26)



Time Series Decomposition

- Suppose, we know that in September electricity consumption was down by 3% from the previous month. Should we conclude that demand for electricity has gone down?
  - No. This might be due to a seasonal effect only. It may be the case that people are using less air conditioners in September than in August.
  - So, we must adjust for the seasonal effect before concluding that there is a long term decline in power consumption.
- Additive Decomposition Model: This model assumes that the time series Y_t is an additive function of three components:
  - $Y_t = Trend_t + Seasonal_t + Irregular_t$
  - Additive model is appropriate if the seasonal fluctuation are same across time periods
- Multiplicative Decomposition Model: This model assumes that the time series Y_t is a product of three components:
  - $Y_t = Trend_t \times Seasonal_t \times Irregular_t$
  - If seasonal fluctuations do vary with the level of the time series because of the trend, then the multiplicative model is to be used

So, let us now continue with that electricity consumption story from the last slide. Suppose we have some data that tells us that in September, electricity consumption was down by 3 percent from the previous month. So, should we conclude that demand for electricity has gone down? Not really because this might be due to seasonal effect only. So, there may be a case that people are using less air conditioning machines in September than in July and August. So, that is what basically I was describing you in that story last time.

So, in a nutshell, what is the learning? So, we must add just for the seasonal effect before concluding that there is a long term decline in power consumption in the area, okay. So, statistical theory tells us that there are two types of decomposition of time series variable available and these two different decomposition models are known as additive decomposition model and the multiplicative decomposition model.

So, in the additive decomposition model, the model assumes that the time series variable Yt is an additive function of three components: these three components are trend, seasonal and irregular. So, you see the equation is written here. Now note down, one interesting fact that here although we have talked about cyclical movements in the time series data we are ignoring the cyclical component in this linear equation for Yt, why? Because just we know we want to avoid the complexity at this point of time because modeling cyclical pattern in an economic variable requires a lot of advanced statistical methods and economic theory. So, we are not getting there.

Now, if we have two alternatives to choose from, one is the additive decomposition model and the other one is the multiplicative decomposition model. Naturally, a question comes to our mind that then when will we go for an additive model? An additive model is appropriate if the seasonal fluctuation are same across different time periods, okay. So later, we will see that we can model this feature by dummy variables.

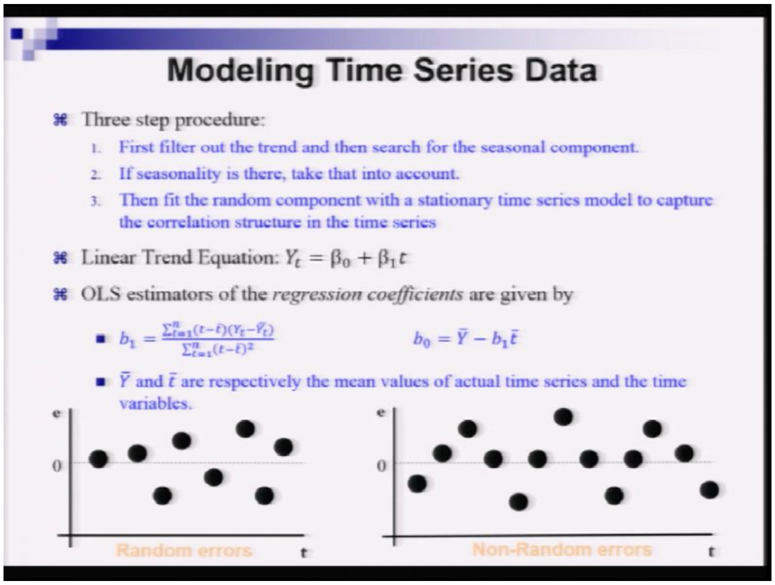
So, now let us move on to the second type of time series, classical time series decomposition model and that is called multiplicative decomposition model. So here, we assume that the time series variable Yt is a product of three components and again, we are ignoring the cyclical component. So, here, Yt is a product of trend seasonal component and the irregular component.

Now again, when shall we use this multiplicative decomposition model over the additive decomposition model? So, if the seasonal fluctuations do vary with the level of time series because of trend trended or some other reason than we can actually apply the multiplicative decomposition model.

So, now, let us assume that we have such a time series data which actually calls for the additive decomposition model and if we assume that additive decomposition model will be perfect for the data at hand then what to do with that data? So, what are the steps that we need to follow if we want to decompose the, this time series data in different components?

And let me tell you why this decomposition actually is required because it may actually challenge your mind that, okay I mean, I have a dataset on y over time, but why do I need to actually decompose it? Actually, the decomposition is required because you finally want to make predictions for the future. That is the ultimate objective of a time series data analysis. So, when you want to predict the future values of y, then you have to model these components like trend seasonal and irregular component very well so that you take care of all sorts of no variation in the data.

(Refer Slide Time: 14:00)



Modeling Time Series Data

- Three step procedure:
  1. First filter out the trend and then search for the seasonal component.
  2. If seasonality is there, take that into account.
  3. Then fit the random component with a stationary time series model to capture the correlation structure in the time series
- Linear Trend Equation: $Y_t = \beta_0 + \beta_1 t$
- OLS estimators of the *regression coefficients* are given by
  - $b_1 = \frac{\sum_{t=1}^{n}(t-\bar{t})(Y_t-\bar{Y}_t)}{\sum_{t=1}^{n}(t-\bar{t})^2}$ $b_0 = \bar{Y} - b_1\bar{t}$
  - $\bar{Y}$ and $\bar{t}$ are respectively the mean values of actual time series and the time variables.

So now, let us look at the three step procedures. So, in the first step, we shall filter out the trend factor and then, we shall search for the seasonal component in the data. And if seasonality is found, then we also have to take into account of that factor. Now, the last step would be when we are done with modeling trended and seasonal components then we have to fit the random component with a stationary time series model to capture the correlation structure in the time series data.

Now, what is stationary time series? That is a complicated discussion and at this moment, I think let us not get into there. So, later in this course, we are going to talk about stationary time series data at length. So, right now, you assume that you have a stationary time series available as the random component of the time series data and let us focus on the trend and the seasonality component only for this one and the next lecture.

So now, let us assume that there is a linear trend that is embedded in our time series variable y. Now, we do not have to assume that there is a linear trend, we can also assume that there is a quadratic taint or cubic trend. But for simplicity, I am showing you a linear trend case and of course, this can be generalized to a quadratic or cubic trend equation.

So here, if I assume that my variable has some hidden trend, then I can, but the variable does not have any seasonal component then I can express these time series variable Yt as a linear equation as beta naught plus beta1 times T and the beta naught is of course, the intercept parameter and the beta1 is basically the slope parameter.

And as you can see that this is matching with what we have learned a couple of lectures before. So, I am here referring our lecture on the linear regression model. So, basically it is a linear equation and we have two population parameters beta naught and beta1 whose values are unknown to us. So, we need to estimate the values for beta naught and beta1.

So, what to do? So, here we can apply the method of ordinary least squares to get the estimates for these unknown population parameters. So, if you apply OLS method, then you can find the OLS versus estimators of the regression coefficients and they are given by the slope coefficient estimate b1 and intercept coefficient estimate b naught.

So, note that, the formula here it seems quite simple to understand, it is basically the same old expression that you have come across in the regression lecture. So, it is basically the sample analogue of the covariance between the time series variable Yt and the time variable t divided by the sample analogue of the time variance variable.

So, here if you concentrate on the formula for b1 now, you see that there are expressions like y bar, t and t bar. What are they? So, y bar t is basically the mean value of the actual time series. So here, you can ignore the subscript t here because you are taking care of the mean over a period of time. So, t can actually fall out because y bar is a constant it is not changing. And then t bar is basically the mean time period.

So, how do you actually throw time in the regression? And let me tell you there are several ways. So, if you have data on say some time periods, like say 20 years or say 24 months, then you can actually name each year or each month by a particular number starting from one. So, it can be like if you are dealing with 20 years data now, the years could be like 1991 to 2010 giving you 20 years of data. But you can actually code 1991 as one, 1992 as two and so on, so forth.

So, ultimately you have this time variable which will take values from one to 20. Now, if you have monthly data for these two years, then the first month say January of 1991, you can code one there in February 1991 you can code two. And similarly, you continue to finally code December of 2010 as the month number 24. And ultimately, you, when you are running the regression you throw these numerical values like 1, 2, 3, 4 to 20 or 2014 the regression to conduct during regression analysis.

So, suppose you code your time period, observation be it month or year by some numbers like, continuous numbers like 1, 2, 3, 4 and then, you run the regression and then you obtain your estimates for the parameters. And once you have that then you can actually make use of these parameter estimates or regression coefficients to have the predicted value for y.

So, for given value of x, as the values of b naught and b1 you can get a yes hat that is the predicted value, and the difference between the actual observation Yt and the predicted value Yt hat that is basically the estimate for the random component in your model and that is basically the residual.

So once, you obtain the residual from the time series regression after fitting the trend line, then what use you can make of that created variable? Now, these residual variables that you just have created is very useful to figure out whether your data set has seasonal component or not. So, there is a very easy graphical approach that you can adopt to check whether the variable that you have just model which is y in this case has a seasonal component embedded in it or not. So, what you have to do? So, you just plot these regression residuals with respect to the time variable that you have used in the regression and see whether there is any pattern or not.

So, for that, let us concentrate on the diagrams that was there in the slide. So here, now, let us focus on the first diagram, which is placed in the southwest corner of the slide. So, here along the x axis I am measuring the time value. So, this is basically the t that goes as the explanatory variable in the linear regression. And this e is basically the residual value that is basically the observed value of y minus the fitted value of y.

And as you probably remember from our discussion on linear regression that residual can be either 0 or positive or negative. So, it can take any value. So, these values are measured along the vertical axis here and you see that I have a hypothetical diagram here for some seven, eight data points. So, here these black circles or spots are basically the residual values corresponding to a particular value of time. And then, you see that there is no pattern in the scatter.

So, we have a broken line that actually measures the zero value of e. So, you see that my data points or the residual data points actually are lying in both sides of these broken lines, some are showing positive residuals and some are showing negative residual values. Overall, there is no pattern. So, if you get this kind of picture or plot, scatter plot then you can decide based on this scatter plot that you are dealing with random errors. So, there is no seasonal component hidden in the data.

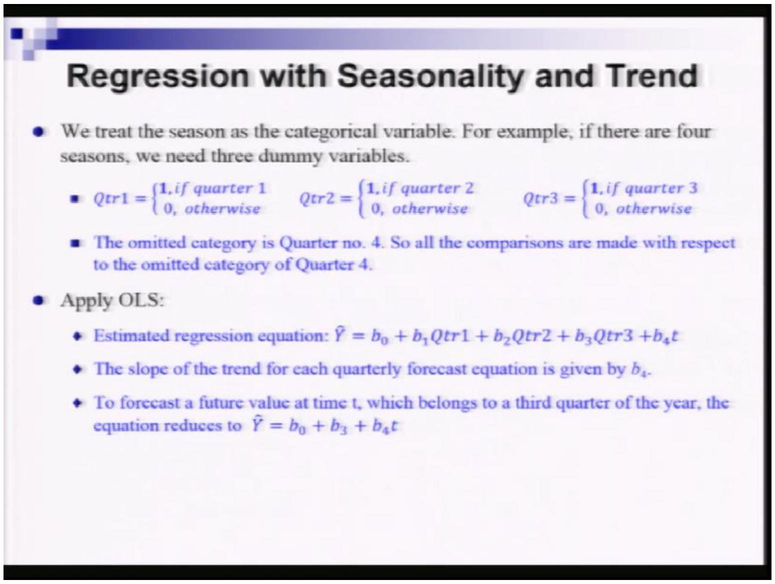
But what if your variable that you just have model, which is yt has a seasonal component hidden in it. So, that is a systematic factor remember and systematic factor is not random. So, if you are not taking care of something very regular or systematic in the variation of y and if you fail to capture that through an explanatory variable in your regression, then the impact of this systematic factor will actually be visible in the residual, because now the residual, which is basically the difference between the observed value and the fitted value will now encompass the hidden variable or the hidden systematic effect that you have not taken care of through regressor in your regression line. And that is what we are going to see next.

So now, you focus on the diagram that is located in the southeast corner of the slide. So, the setup is same. So, you have the residuals plotted and here, you see these black dots are showing some kind of trend. So, there are, first three dots if you see, they are increasing. Then you see there are two dots which are falling then again next two dots, it is increasing and some dots are lying on the broken line, which is basically showing 0 value.

So, there is a pattern it is mostly like a sine, cosine kind of fluctuation that you are observing. Although you are not observing a curve, here you are observing discrete points. But some kind of sine cosine type of curve is there, okay that is evident from the scatter plot here.

So, if you observe this kind of scatter plot which shows some kind of pattern in the residual and the pattern is something like a sine cosine curve, then you can be sure that your variable time series variable Yt actually has seasonal component and you have forgotten to model that.

(Refer Slide Time: 25:14)



Regression with Seasonality and Trend

- We treat the season as the categorical variable. For example, if there are four seasons, we need three dummy variables.
  - $Qtr1 = \begin{cases} 1, if\ quarter\ 1 \\ 0,\ otherwise \end{cases}$ $Qtr2 = \begin{cases} 1, if\ quarter\ 2 \\ 0,\ otherwise \end{cases}$ $Qtr3 = \begin{cases} 1, if\ quarter\ 3 \\ 0,\ otherwise \end{cases}$
  - The omitted category is Quarter no. 4. So all the comparisons are made with respect to the omitted category of Quarter 4.
- Apply OLS:
  - Estimated regression equation: $\hat{Y} = b_0 + b_1 Qtr1 + b_2 Qtr2 + b_3 Qtr3 + b_4 t$
  - The slope of the trend for each quarterly forecast equation is given by b_4.
  - To forecast a future value at time t, which belongs to a third quarter of the year, the equation reduces to $\hat{Y} = b_0 + b_3 + b_4 t$

Okay. So, let us assume that you first tried linear trend regression with your Yt variable and then, you examine the scatter plot of the residuals that is generated from that linear trend regression and you found that there are sample evidences, which are saying that there could be seasonality impact embedded in the Yt variable, which you have modeled. So, in that case what to do?

So, if you have some evidence, statistical evidence for the presence of seasonality then there are many ways you can take care of it. And today, I am going to only talk about one method and that is basically the method of dummy variable regression. So, in the context of ANCOVA, we have already studied the dummy variable. So, I am not defining dummy variable again.

So, if you have forgotten the definition of dummy variable, so, let me know spend couple of seconds. So, a dummy variable basically is an indicator variable, which takes value 1 when a qualitative variable takes a particular category or level and it takes 0 value if the quality variable does not take that particular level or category.

So here, in this case of time series data analysis, how do I make use of dummy variable regression technique? So, here we treat seasons as the categorical variable, okay. So, if we have four different seasons in a year, namely summer, autumn, winter and spring, so, in our country the economic calendar is kind of linked with the seasons and so, we can make use of the quarter to define different seasons. So, if we do so, then we have four quarters in a year. So, to avoid the dummy variable trap, we have to make use of three levels of these qualitative variable season. And then, we can actually define three dummy variables.

So, let us see, which way we can define these dummies? Well, we have to fix a base and then with respect to that base we have to define three dummy variables in this case. So, if you remember the discussion in the context of ANCOVA, I said that if there are g levels for a qualitative variable, then we can at max define g minus one dummy variables.

So, similarly here, as season is the qualitative variable, here we can actually introduce three dummy variables because we are assuming that our season qualitative variable has four levels. So here, we are defining three dummy variables, we are dealing with four quarters and we are omitting the category or level quarter number four. So, all three other dummy variables are defined with respect to the base category, quarter four.

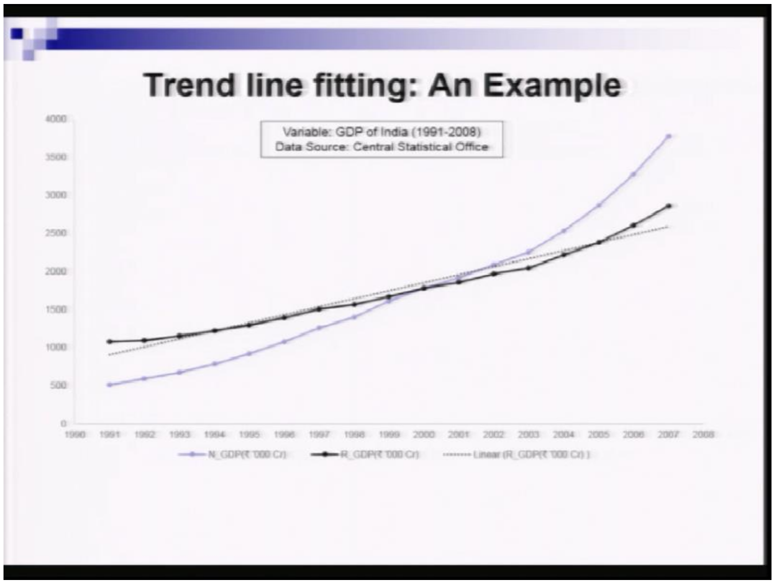
So, here are my three dummy variables. So, first one is quarter one, so, that takes value one. If we have an observation from quarter one, the second dummy variable is quarter two, and that takes value one, if we have an observation from quarter two. And finally, we have dummy variable quarter three, if we find an observation coming from quarter three of a particular year, then we give value one to it. Otherwise, it takes 0 value.

So, once you have defined the dummy variables, then you have actually four variables, okay, one is the continuous variable that is basically the trend variable. And then, you have three quarter dummies. So, again you see it is kind of an ANCOVA. So, here although we are not doing ANCOVA, we can still apply the ordinary least square method. And here I am showing you the estimated regression equation to be y hat equals to B naught that is the intercept parameter estimate, then b1 is basically the regression coefficient or estimate for the population parameter that is attached to the quarter one dummy variable. And similarly, you can interpret what are b2, b3 and b4.

So now you that that slope of that trend variable t is basically same for all quarter, okay and that is given by b4. Now, if you want to forecast a future value at time t that belongs to say third quarter of the year then, how do you proceed? So, then the equation the estimated regression equation reduces to y hat equals to B naught plus b3 plus b4 times t, why?

So, where are my coefficients b1 and b2? They will fall out from the originally estimated regression equation because as the observation is coming from the third quarter of a particular year, the dummy variables quarter one and quarter two will take zero values for this case and hence, b1 and b2 will not show up in the reduced equation. So, we have this simple equation now, here you plug the value of t and you get the future value at time t.

(Refer Slide Time: 31:16)



Trend line fitting: An Example

Now, we are going to end today’s lecture by looking at a very simple linear trend fitting exercise based on Indian GDP data. Let me assume that I have data on GDP of India and I have data from 1991 to 2008. So, that is basically my time series variable and Data Sources Central Statistical Office Government of India. So here, along the x axis I am measuring the years, I am showing different values of years and then along the y axis I am measuring the GDP.

Now, note that when you get this data, you actually have two different series that is reported by the Government one is the nominal GDP and the other one is real GDP. So, real GDP means that GDP is expressed in constant price of a particular year. Now, when you are trying to conduct a trend analysis, which one to use? You should make use of the real GDP when price level is fixed for a particular year.

Why? Because I told you in the previous two lectures that there is inflationary pressure when you are looking at value of an economic variable and when you have data for say more than two, three years inflation is varying and that is why two nominal values over two different time periods, it is not ideal to compare them and do any statistical analysis involving them.

So, you have to actually make use of the GDP deflator to convert the nominal values of the GDP and then, you get the real GDP and then you can try to fit a trend line and then get the estimator trended equation that is precisely we are doing here. So, let us go back to the diagram again. Here, the n dash GDP actually measures the nominal GDP in rupees 1,000 crores and there is GDP deflator data available from the Reserve Bank of India and then, we can make use of that to convert the nominal GDP into the real GDP.

And here, the index that we are using that keeps the price level constant at 1999, 2000 level. And for that particular year, the deflator takes value 100 and for other years it assumes different values. Now, note that, when we have the real data, we can actually feed linear trend equation and if we apply OLS to this real GDP data, then the fitted straight line is given by this broken line, this dotted line. So, let us stop here for the time being. I will be back with more discussion on time series data analysis in the next lecture. See you then. Thank you.