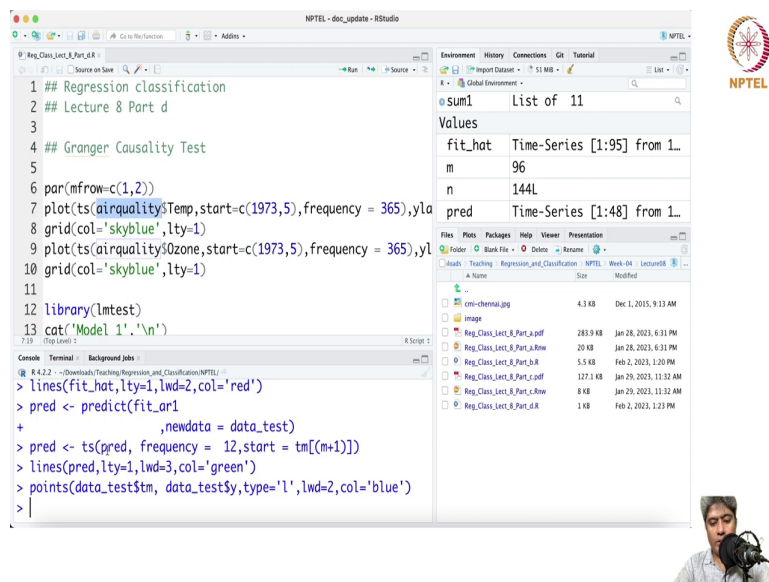


Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 29
Hands on with R Part - 7

Hello all. Welcome back to the last part of video lecture series of 8. Now, we are going to do the Hands-on for Granger causality test.

(Refer Slide Time: 00:29)



The screenshot displays the RStudio interface. The main editor shows R code for regression classification and Granger causality test. The environment pane on the right shows the 'sum1' list of 11 values, including 'fit_hat', 'm', 'n', and 'pred'. The file explorer on the right shows the current directory structure.

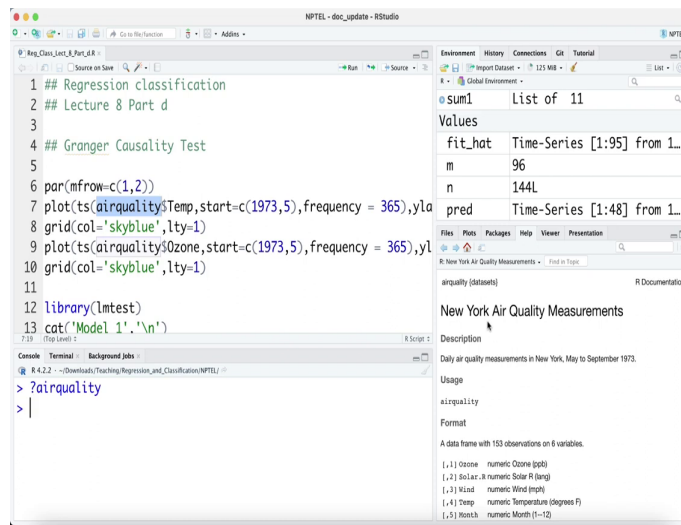
```
1 ## Regression classification
2 ## Lecture 8 Part d
3
4 ## Granger Causality Test
5
6 par(mfrow=c(1,2))
7 plot(ts(airquality$Temp,start=c(1973,5),frequency = 365),ylab="Temp",lty=1)
8 grid(col="skyblue",lty=1)
9 plot(ts(airquality$Ozone,start=c(1973,5),frequency = 365),ylab="Ozone",lty=1)
10 grid(col="skyblue",lty=1)
11
12 library(lmtest)
13 cat("Model 1'."\n")
14
15 > lines(fit_hat,lty=1,lwd=2,col='red')
16 > pred <- predict(fit_ar1,
17                 ,newdata = data_test)
18 > pred <- ts(pred, frequency = 12,start = tm[(m+1)])
19 > lines(pred,lty=1,lwd=3,col='green')
20 > points(data_test$tm, data_test$y,type='l',lwd=2,col='blue')
21 >
```

Variable	Value
sum1	List of 11
fit_hat	Time-Series [1:95] from 1...
m	96
n	144L
pred	Time-Series [1:48] from 1...

Name	Size	Modified
...
con-chemical.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
image
Reg_Class_Lect_8_Part_a.pdf	283.9 KB	Jan 28, 2023, 6:31 PM
Reg_Class_Lect_8_Part_a.Rnw	20 KB	Jan 28, 2023, 6:31 PM
Reg_Class_Lect_8_Part_b.R	5.5 KB	Feb 2, 2023, 1:20 PM
Reg_Class_Lect_8_Part_c.pdf	127.1 KB	Jan 29, 2023, 11:32 AM
Reg_Class_Lect_8_Part_c.Rnw	8 KB	Jan 29, 2023, 11:32 AM
Reg_Class_Lect_8_Part_d.R	1 KB	Feb 2, 2023, 1:23 PM

So, first we will consider this airquality dataset.

(Refer Slide Time: 00:45)



The screenshot shows the RStudio interface with the following R code in the editor:

```
1 ## Regression classification
2 ## Lecture 8 Part d
3
4 ## Granger Causality Test
5
6 par(mfrow=c(1,2))
7 plot(ts(airquality$Temp,start=c(1973,5),frequency = 365),yla
8 grid(col='skyblue',lty=1)
9 plot(ts(airquality$Ozone,start=c(1973,5),frequency = 365),yl
10 grid(col='skyblue',lty=1)
11
12 library(lmtest)
13 cat('Model 1'. '\n')
```

The console shows the command `> ?airquality` and the output:

```
> |
> |
```

The Environment pane shows the following variables:

sum1	List of 11
fit_hat	Time-Series [1:95] from 1...
m	96
n	144L
pred	Time-Series [1:48] from 1...

The Files pane shows the following files:

- R: New York Air Quality Measurements
- airquality (dataset)
- New York Air Quality Measurements

The Description pane shows the following information:

Description
Daily air quality measurements in New York, May to September 1973.

Usage
airquality

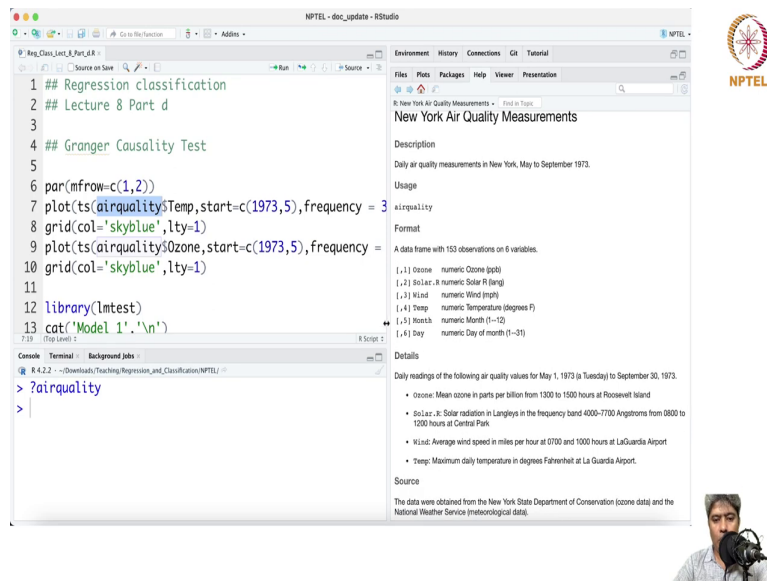
Format
A data frame with 153 observations on 6 variables.

[,1]	Ozone	numeric: Ozone (ppb)
[,2]	Solar.R	numeric: Solar R (lang)
[,3]	Wind	numeric: Wind (mph)
[,4]	Temp	numeric: Temperature (degrees F)
[,5]	Month	numeric: Month (1--12)



Say this test comes with the basic datasets package.

(Refer Slide Time: 00:51)



The screenshot displays the RStudio interface. The editor window contains the following R code:

```
1 ## Regression classification
2 ## Lecture 8 Part d
3
4 ## Granger Causality Test
5
6 par(mfrow=c(1,2))
7 plot(ts(airquality$Temp,start=c(1973,5),frequency = 3
8 grid(col='skyblue',lty=1)
9 plot(ts(airquality$Ozone,start=c(1973,5),frequency =
10 grid(col='skyblue',lty=1)
11
12 library(lmtest)
13 cat('Model 1'. '\n')
```

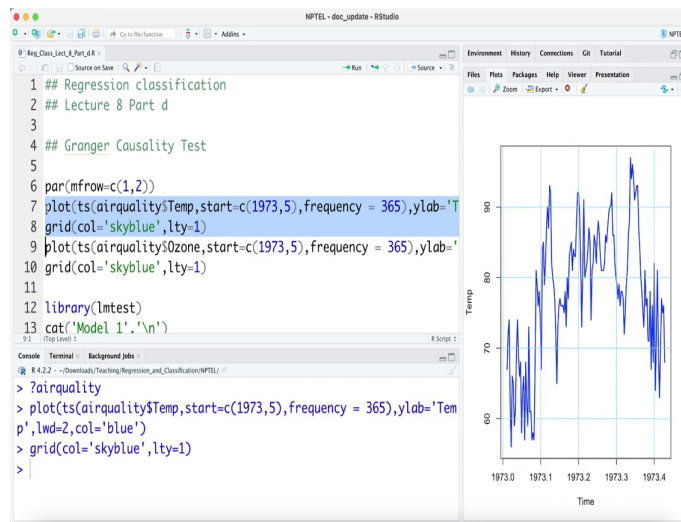
The console shows the command `> ?airquality` and the help page for the `airquality` dataset. The help page includes the following information:

- Description:** Daily air quality measurements in New York, May to September 1973.
- Usage:** `airquality`
- Format:** A data frame with 153 observations on 6 variables.
 - [,1] Ozone: numeric Ozone (ppb)
 - [,2] Solar.R: numeric Solar R (lang)
 - [,3] Wind: numeric Wind (mph)
 - [,4] Temp: numeric Temperature (degrees F)
 - [,5] Month: numeric Month (1-12)
 - [,6] Day: numeric Day of month (1-31)
- Details:** Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.
 - Ozone:** Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
 - Solar.R:** Solar radiation in Langley's in the frequency band 4000-7700 Angstroms from 0800 to 1200 hours at Central Park
 - Wind:** Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
 - Temp:** Maximum daily temperatures in degrees Fahrenheit at La Guardia Airport.
- Source:** The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data).

And in the air, it is if you look into the description of the dataset its New York Air Quality Measurement. So, Daily Air Quality Measurements in New York between May and September of 1973. So, data its a data frame it has 153 observation on 6 variables ozone, solar, wind temperature, month and day.

Now, we are only focusing on temperature and ozone, but you can try with the other variable as well. So, the first thing we are going to plot is temperature dataset.

(Refer Slide Time: 01:42)



The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
1 ## Regression classification
2 ## Lecture 8 Part d
3
4 ## Granger Causality Test
5
6 par(mfrow=c(1,2))
7 plot(ts(airquality$Temp,start=c(1973,5),frequency = 365),ylab='T
8 grid(col='skyblue',lty=1)
9 plot(ts(airquality$Ozone,start=c(1973,5),frequency = 365),ylab='
10 grid(col='skyblue',lty=1)
11
12 library(lmtest)
13 cat('Model 1'. '\n')
```

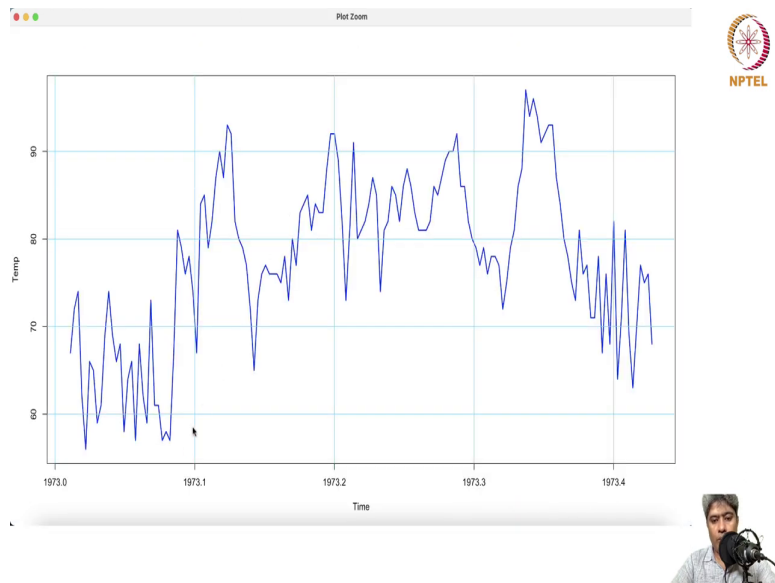
The console on the bottom left shows the execution of the following commands:

```
> ?airquality
> plot(ts(airquality$Temp,start=c(1973,5),frequency = 365),ylab='Tem
p',lwd=2,col='blue')
> grid(col='skyblue',lty=1)
>
```

The plot window on the right shows a time series plot of temperature data. The y-axis is labeled 'Temp' and ranges from 60 to 90. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot displays a blue line representing the temperature data, with a light blue grid overlaid. The data shows a clear seasonal pattern with peaks around 1973.1 and 1973.3, and troughs around 1973.0 and 1973.2.

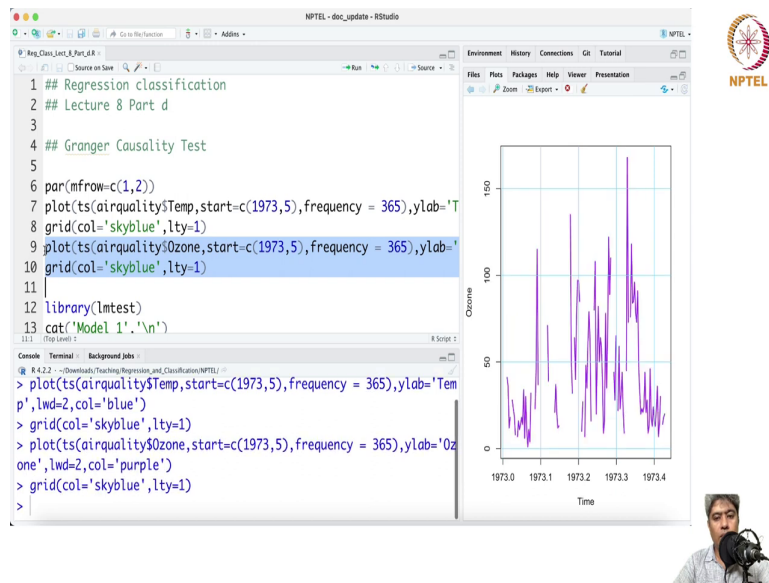


(Refer Slide Time: 01:46)



That is how like temperature dataset has.

(Refer Slide Time: 01:52)



The screenshot displays an RStudio interface. The main editor window contains the following R code:

```
1 ## Regression classification
2 ## Lecture 8 Part d
3
4 ## Granger Causality Test
5
6 par(mfrow=c(1,2))
7 plot(ts(airquality$Temp,start=c(1973,5),frequency = 365),ylab='T
8 grid(col='skyblue',lty=1)
9 plot(ts(airquality$Ozone,start=c(1973,5),frequency = 365),ylab='
10 grid(col='skyblue',lty=1)
11
12 library(lmtest)
13 cat('Model 1'. '\n')
```

The console window shows the execution of the following commands:

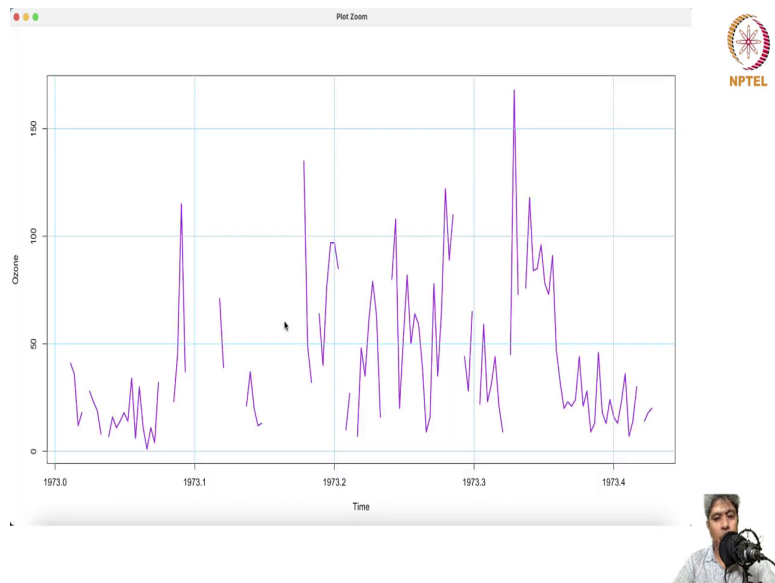
```
> plot(ts(airquality$Temp,start=c(1973,5),frequency = 365),ylab='Temp',lwd=2,col='blue')
> grid(col='skyblue',lty=1)
> plot(ts(airquality$Ozone,start=c(1973,5),frequency = 365),ylab='Ozone',lwd=2,col='purple')
> grid(col='skyblue',lty=1)
>
```

The plot window displays a time-series plot of ozone levels. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The plot shows a highly volatile time series with a clear seasonal pattern, characterized by regular peaks and troughs. The data points are plotted in purple, and the plot area is enclosed by a light blue grid.

The NPTEL logo is visible in the top right corner of the RStudio window.

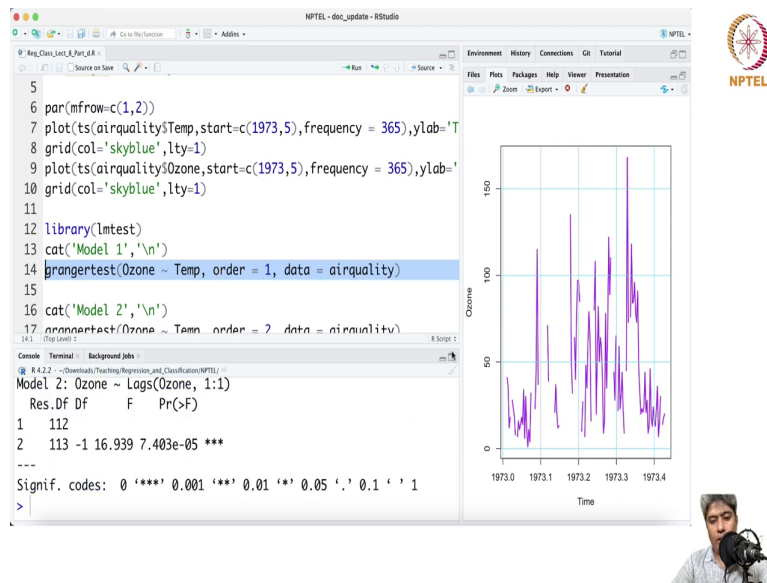
This is the ozone dataset as we have seen.

(Refer Slide Time: 01:55)



Looks like there are some missing values are also there in the ozone time series.

(Refer Slide Time: 02:00)





The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
5
6 par(mfrow=c(1,2))
7 plot(ts(airquality$Temp,start=c(1973,5),frequency = 365),ylab='T
8 grid(col='skyblue',lty=1)
9 plot(ts(airquality$Ozone,start=c(1973,5),frequency = 365),ylab='
10 grid(col='skyblue',lty=1)
11
12 library(lmtest)
13 cat('Model 1',"\n")
14 grangertest(Ozone ~ Temp, order = 1, data = airquality)
15
16 cat('Model 2',"\n")
17 grangertest(Ozone ~ Temp, order = 2, data = airquality)
```

The console output shows the results for Model 2:

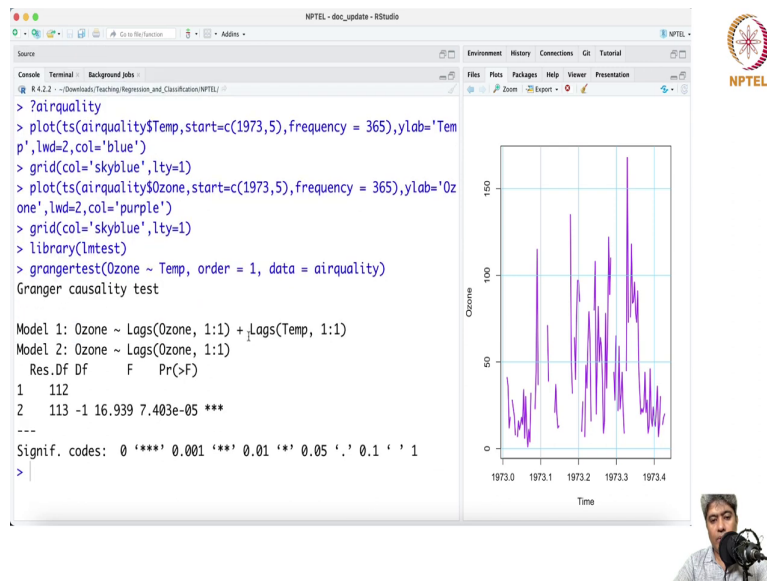
```
Model 2: Ozone ~ Lags(Ozone, 1:1)
Res.Df Df F Pr(>F)
1 112
2 113 -1 16.939 7.403e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The plot window on the right shows a time-series plot of Ozone concentration over time. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with a clear seasonal pattern, with peaks around 1973.1 and 1973.3, and troughs around 1973.0 and 1973.2. A light blue grid is visible in the background of the plot.



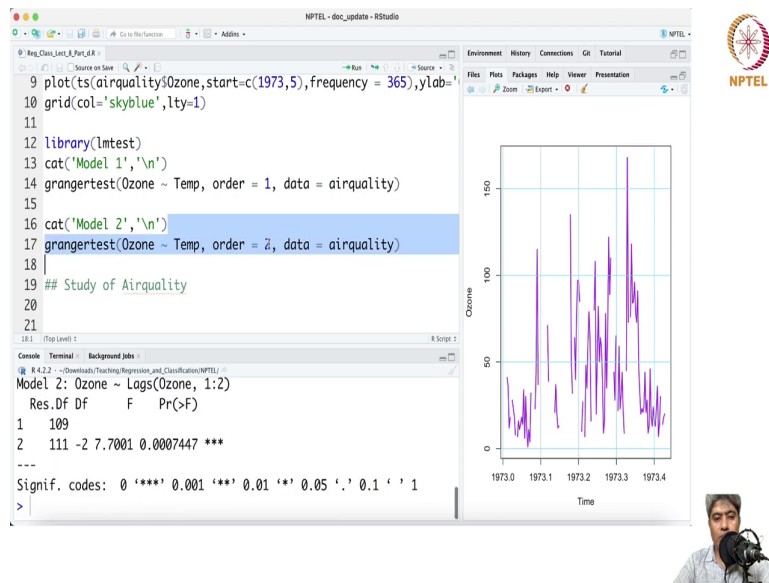
So, the first thing we are running here is in from the LM test. The first model that we are trying to fit is the grangertest where Ozone is a function of Temperature and with only lag 1 and if we run this.

(Refer Slide Time: 02:25)



And data is over air quality we run this the P value this is the F test run. So, the Model 1 is Ozone with Lag 1 Ozone and Lag 1 Temperature and then the second model is only the null model and then it did a test whether the lag model has effect or not and then it says that ok the null model is not right.

(Refer Slide Time: 02:52)





The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
9 plot(ts(airquality$Ozone,start=c(1973,5),frequency = 365),ylab='Ozone')
10 grid(col='skyblue',lty=1)
11
12 library(lmtest)
13 cat('Model 1', '\n')
14 grangertest(Ozone ~ Temp, order = 1, data = airquality)
15
16 cat('Model 2', '\n')
17 grangertest(Ozone ~ Temp, order = 2, data = airquality)
18
19 ## Study of Airquality
20
21
```

The console output shows the results for Model 2:

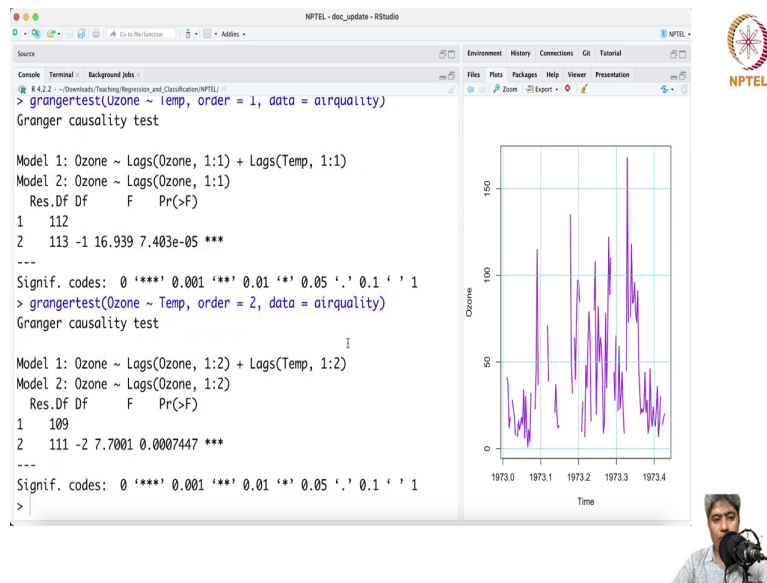
```
Model 2: Ozone ~ Lags(Ozone, 1:2)
Res.Df Df F Pr(>F)
1 109
2 111 -2 7.7001 0.0007447 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The plot on the right shows a time series of Ozone levels from 1973.0 to 1973.4. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with a clear upward trend and a seasonal pattern. A light blue grid is overlaid on the plot.



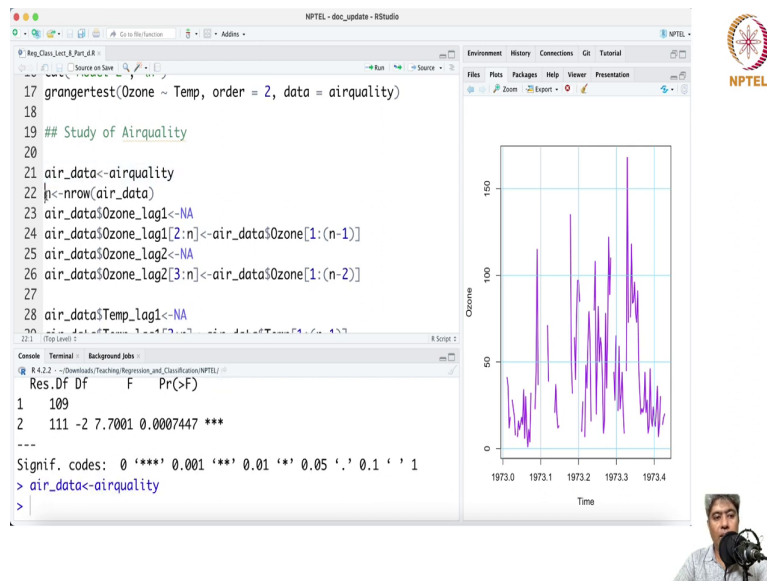
So, that means, temperature does have a effect on the Ozone. Similarly, we run the second model and in this ozone again as a function of Temperature, but now we are going up to the second order of the granger causal test model.

(Refer Slide Time: 03:15)



Its essentially autoregressive model essentially its autoregressive model and P value is still small. So, we can say that temperature does have effect on ozone.

(Refer Slide Time: 03:31)





The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
17 grangertest(Ozone ~ Temp, order = 2, data = airquality)
18
19 ## Study of Airquality
20
21 air_data<-airquality
22 h<-nrow(air_data)
23 air_data$Ozone_lag1<-NA
24 air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
25 air_data$Ozone_lag2<-NA
26 air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
27
28 air_data$Temp_lag1<-NA
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30
```

The console on the left shows the output of the `grangertest` function:

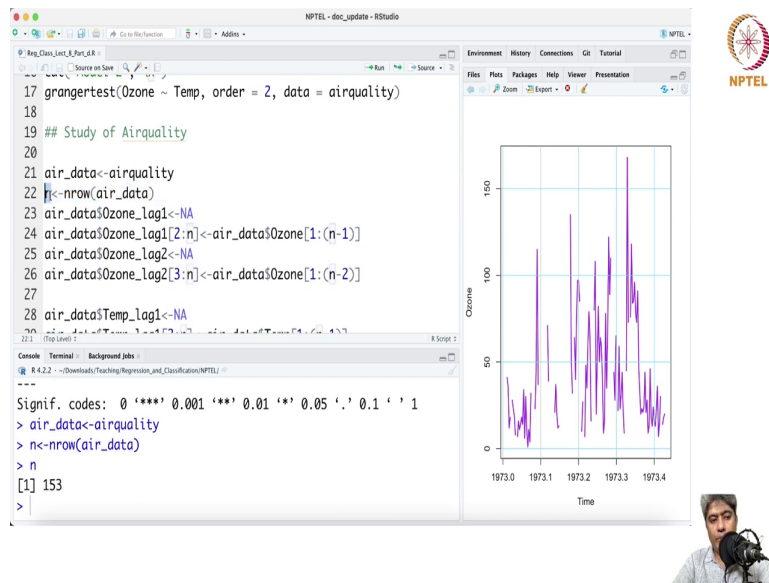
```
Res.Df Df F Pr(>F)
1 109
2 111 -2 7.7001 0.0007447 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> air_data<-airquality
>
```

The plot window on the right shows a time series plot of Ozone concentration over time. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks, notably around 1973.1 and 1973.3.



So, in order to understand a little bit more how this whole thing is working. So, we take the air quality data.

(Refer Slide Time: 03:43)



```
17 grangertest(Ozone ~ Temp, order = 2, data = airquality)
18
19 ## Study of Airquality
20
21 air_data<-airquality
22 n<-nrow(air_data)
23 air_data$Ozone_lag1<-NA
24 air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
25 air_data$Ozone_lag2<-NA
26 air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
27
28 air_data$Temp_lag1<-NA
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30
31
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> air_data<-airquality
> n<-nrow(air_data)
> n
[1] 153
>
```

The plot shows Ozone levels on the y-axis (ranging from 0 to 150) and Time on the x-axis (ranging from 1973.0 to 1973.4). The data is represented by a purple line with vertical error bars, showing a clear seasonal pattern with peaks around 1973.1 and 1973.3, and troughs around 1973.0 and 1973.2.

n is the number of samples. First what we and now if you look into the let us look into the `air_data`.

(Refer Slide Time: 03:53)

The screenshot shows an RStudio session with the following code in the script editor:

```
17 grangertest(Ozone ~ Temp, order = 2, data = airquality)
18
19 ## Study of Airquality
20
21 air_data<-airquality
22 n<-nrow(air_data)
23 air_data$Ozone_lag1<-NA
24 air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
25 air_data$Ozone_lag2<-NA
26 air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
27
28 air_data$Temp_lag1<-NA
```

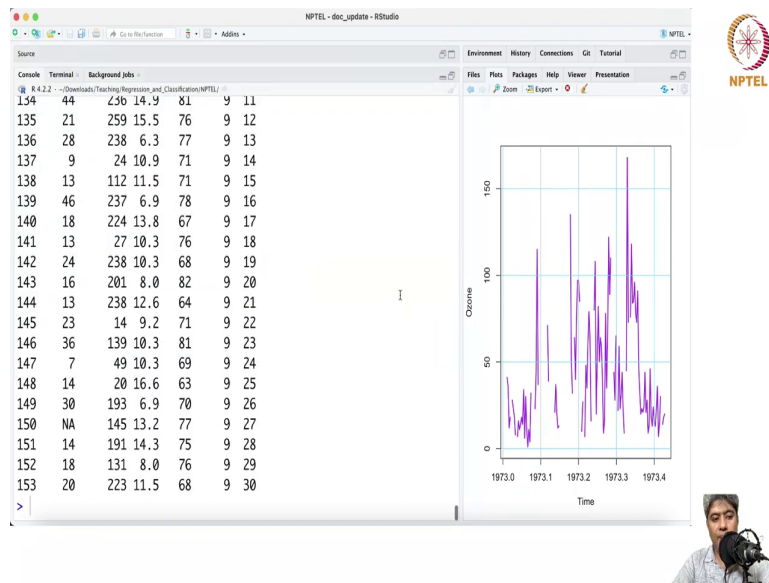
The console output shows the following data:

Time	Ozone	Temp	Wind	Humid	Windspeed	
148	14	20	16.6	63	9	25
149	30	193	6.9	70	9	26
150	NA	145	13.2	77	9	27
151	14	191	14.3	75	9	28
152	18	131	8.0	76	9	29
153	20	223	11.5	68	9	30

The plot on the right shows Ozone levels over time from 1973.0 to 1973.4. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks reaching above 100.

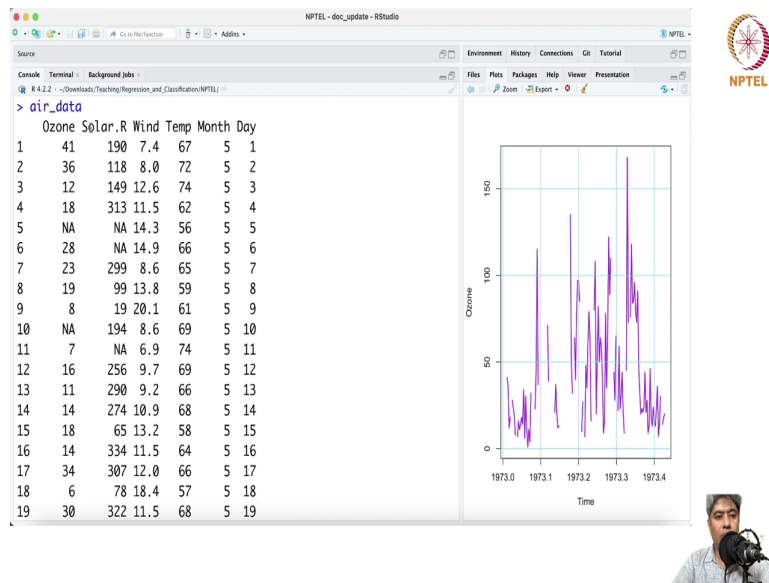


(Refer Slide Time: 03:56)



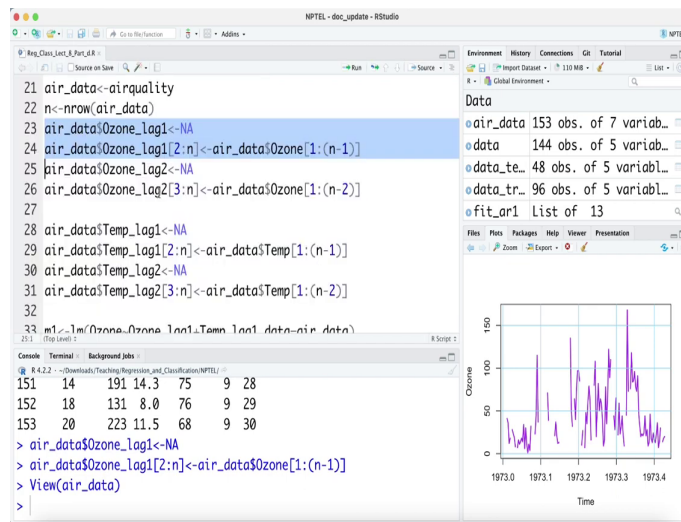
So, that is how the data looks like, ok.

(Refer Slide Time: 04:01)



That is how the Ozone Solar Solar.R radiation Wind Solar.R stands for radiation Wind Temperature Month and Day. Now, there are some NA observations are available.

(Refer Slide Time: 04:25)



The screenshot shows the RStudio interface with the following R code in the script editor:

```
21 air_data<-airquality
22 n<-nrow(air_data)
23 air_data$Ozone_lag1<-NA
24 air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
25 air_data$Ozone_lag2<-NA
26 air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
27
28 air_data$Temp_lag1<-NA
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30 air_data$Temp_lag2<-NA
31 air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
32
33 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
```

The console shows the output of the first three lines of code:

```
151 14      191 14.3  75    9 28
152 18      131  8.0  76    9 29
153 20      223 11.5  68    9 30
```

The environment pane shows the following objects:

- air_data: 153 obs. of 7 variables
- data: 144 obs. of 5 variables
- data_te: 48 obs. of 5 variables
- data_tr: 96 obs. of 5 variables
- fit_ar1: List of 13

The plot pane shows a line graph of Ozone concentration over time (1973.0 to 1973.4). The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks exceeding 100.



So, it is always bit difficult how you do this you know imputation, but for the time being we are not handling the missing data. We are only going to use the data which is fully available to us, but for the time being. So, we suppose we want to fit the first model we create the lag data set ok.

(Refer Slide Time: 04:47)

The screenshot shows the RStudio interface with the following components:

- Environment:** Shows a list of objects: `air_data` (153 obs. of 7 variab...), `data` (144 obs. of 5 variab...), `data_te...` (48 obs. of 5 variabl...), `data_tr...` (96 obs. of 5 variabl...), and `fit_ar1` (List of 13).
- Table:** A data frame with 16 rows and 7 columns: `Ozone`, `Solar.R`, `Wind`, `Temp`, `Month`, `Day`, and `Ozone_lag1`. The first few rows are:

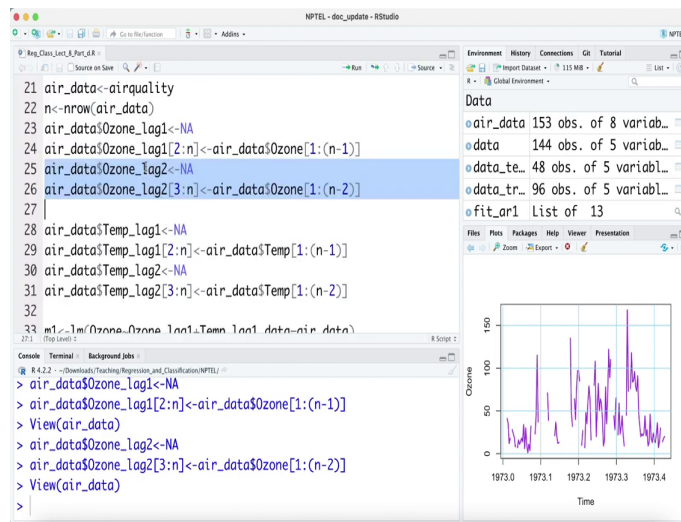
	Ozone	Solar.R	Wind	Temp	Month	Day	Ozone_lag1
1	41	190	7.4	67	5	1	NA
2	36	118	8.0	72	5	2	41
3	12	149	12.6	74	5	3	36
4	18	313	11.5	62	5	4	12
5	NA	NA	14.3	56	5	5	18
- Console:** Shows the following R code and output:

```
> air_data$Ozone_lag1<-NA
> air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
> View(air_data)
>
```
- Plot:** A line plot of Ozone concentration over time (1973.0 to 1973.4). The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time'.



So, here we have created the lag. So, 41 was here we just created lag 36 just brought it down by 1.

(Refer Slide Time: 04:58)



The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
21 air_data<-airquality
22 n<-nrow(air_data)
23 air_data$Ozone_lag1<-NA
24 air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
25 air_data$Ozone_lag2<-NA
26 air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
27
28 air_data$Temp_lag1<-NA
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30 air_data$Temp_lag2<-NA
31 air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
32
33 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
```

The console on the bottom left shows the execution of the first few lines of code:

```
> air_data$Ozone_lag1<-NA
> air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
> View(air_data)
> air_data$Ozone_lag2<-NA
> air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
> View(air_data)
```

The Environment pane on the right shows the objects created in the workspace:

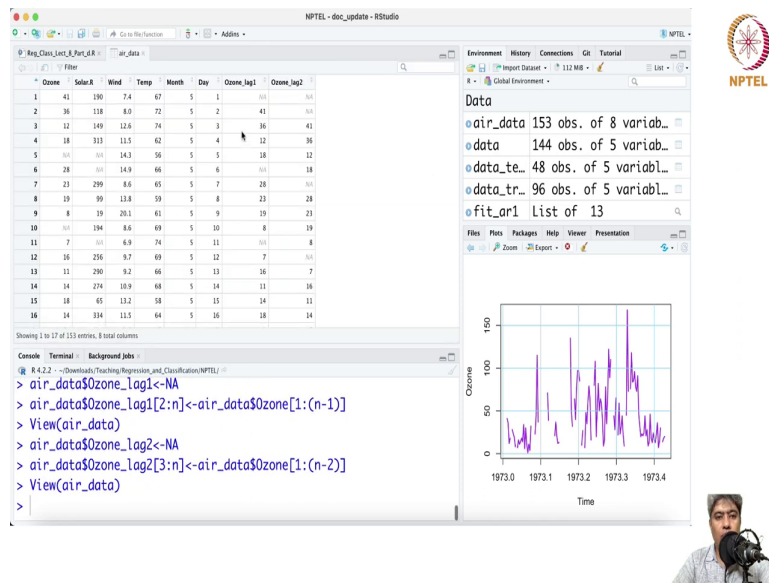
- air_data: 153 obs. of 8 variab...
- data: 144 obs. of 5 variab...
- data_te: 48 obs. of 5 variab...
- data_tr: 96 obs. of 5 variab...
- fit_ar1: List of 13

A time series plot of Ozone concentration is displayed in the bottom right pane. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks exceeding 100.



Then we created the lag2 variable.



(Refer Slide Time: 05:00)



The screenshot displays the RStudio interface. The main window shows a data frame with columns: Ozone, Solar.R, Wind, Temp, Month, Day, Ozone_lag1, and Ozone_lag2. The console shows the following R commands and their output:

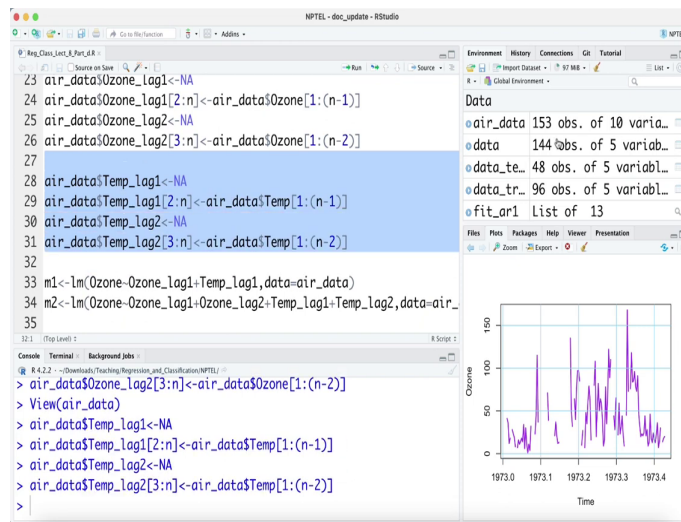
```
R 4.2.2 -> Downloads/Teaching/Regression_and_Classification/NPTEL/
> air_data$Ozone_lag1<-NA
> air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
> View(air_data)
> air_data$Ozone_lag2<-NA
> air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
> View(air_data)
>
```

The right-hand pane shows a list of data objects, including 'air_data' (153 obs. of 8 variables) and 'fit_ar1' (List of 13). Below the list is a time-series plot of 'Ozone' levels from 1973.0 to 1973.4, showing a clear seasonal pattern with peaks around 1973.1 and 1973.3.



So, now you have see the lag2 variables have been created ok.

(Refer Slide Time: 05:07)



The screenshot shows the RStudio interface with the following R code in the script editor:

```
23 air_data$Ozone_lag1<-NA
24 air_data$Ozone_lag1[2:n]<-air_data$Ozone[1:(n-1)]
25 air_data$Ozone_lag2<-NA
26 air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
27
28 air_data$Temp_lag1<-NA
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30 air_data$Temp_lag2<-NA
31 air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
32
33 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
34 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_
35
```

The console shows the execution of the following commands:

```
> air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
> View(air_data)
> air_data$Temp_lag1<-NA
> air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
> air_data$Temp_lag2<-NA
> air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
>
```

The Environment pane on the right shows the following objects:

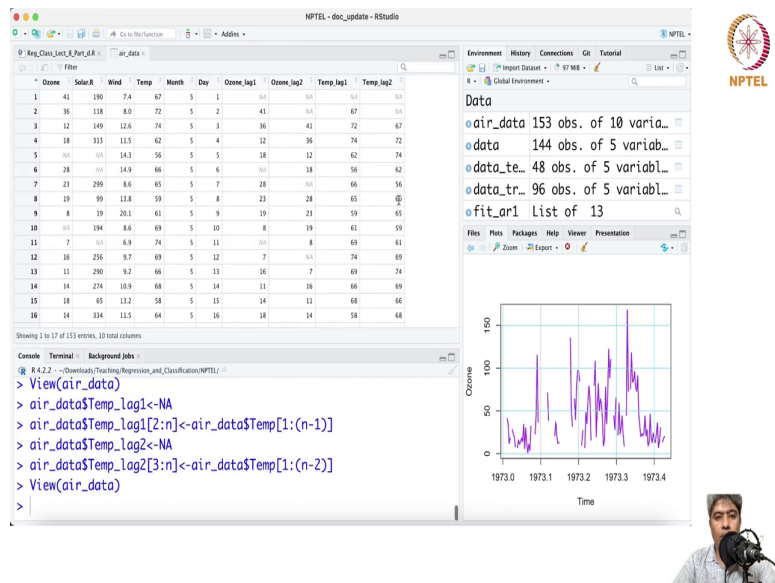
- air_data: 153 obs. of 10 variab...
- data: 144 obs. of 5 variab...
- data_te: 48 obs. of 5 variabl...
- data_tr: 96 obs. of 5 variabl...
- fit_ar1: List of 13

A plot of Ozone vs Time is displayed, showing a time series of ozone levels from 1973.0 to 1973.4. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks reaching above 100.



Then similarly we create the Temperature_lag1 and Temperature_lag2 variables.

(Refer Slide Time: 05:12)



The screenshot shows the RStudio interface with the following components:

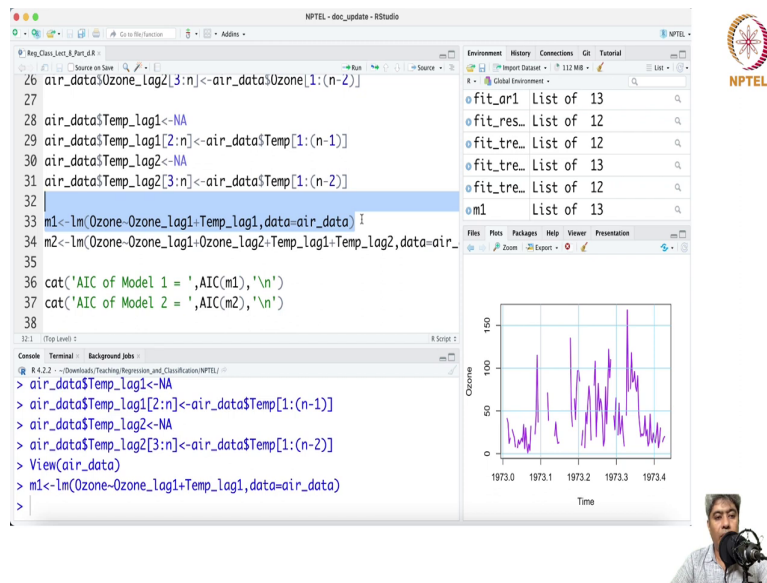
- Environment:** Lists data objects: `air_data` (153 obs. of 10 variables), `data` (144 obs. of 5 variables), `data_te` (48 obs. of 5 variables), `data_tr` (96 obs. of 5 variables), and `fit_ar1` (List of 13).
- Files:** Shows a plot of Ozone vs Time.
- Console:** Contains the following R code:

```
> View(air_data)
> air_data$Temp_lag1<-NA
> air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
> air_data$Temp_lag2<-NA
> air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
> View(air_data)
>
```

The plot shows Ozone concentration (Y-axis, 0 to 150) over Time (X-axis, 1973.0 to 1973.4). The data shows a clear seasonal pattern with peaks around 1973.1 and 1973.3, and troughs around 1973.0 and 1973.2.

So, Temperature_lag1, Temperature_lag2 variable have been created.

(Refer Slide Time: 05:19)



The screenshot displays the RStudio interface. The script editor contains the following R code:

```
26 air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
27
28 air_data$Temp_lag1<-NA
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30 air_data$Temp_lag2<-NA
31 air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
32
33 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
34 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_data)
35
36 cat('AIC of Model 1 = ',AIC(m1),'\n')
37 cat('AIC of Model 2 = ',AIC(m2),'\n')
38
```

The Environment pane on the right shows the following objects:

- fit_ar1 List of 13
- fit_res_ List of 12
- fit_tre_ List of 12
- fit_tre_ List of 13
- fit_tre_ List of 12
- m1 List of 13

The Console pane shows the execution of the following commands:

```
> air_data$Temp_lag1<-NA
> air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
> air_data$Temp_lag2<-NA
> air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
> View(air_data)
> m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
>
```

The Plots pane displays a line graph of Ozone concentration over time. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks reaching approximately 150.

The NPTEL logo is visible in the top right corner of the RStudio window.

Now, if you fit the model with Ozone_lag1 and Temperature_ag1 ok.

(Refer Slide Time: 05:36)

The screenshot shows the RStudio interface with the following R code in the script editor:

```
27
28 air_data$Temp_lag1<-NA
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30 air_data$Temp_lag2<-NA
31 air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
32
33 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
34 m10 <-lm(Ozone~Ozone_lag1,data=air_data)
35 anova(m10,m1)
36
37 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_
38
39 cat('\nATC of Model 1 = ', ATC(m1), '\n')
```

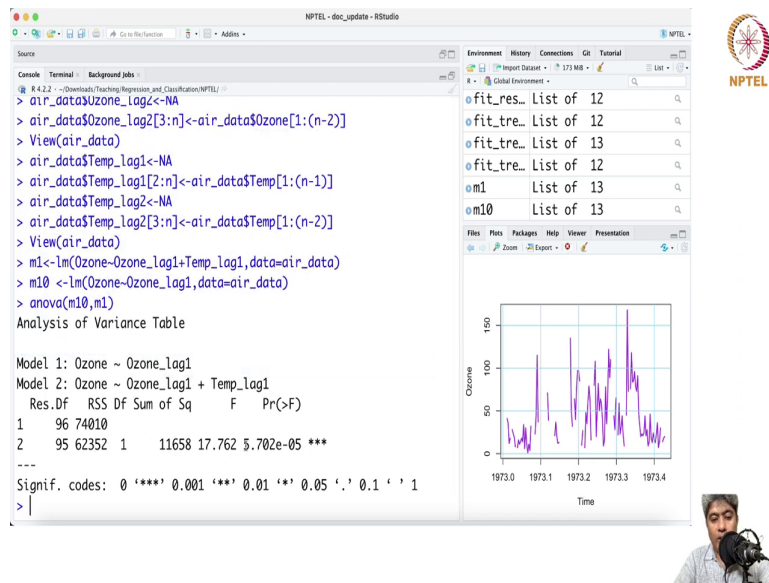
The console shows the execution of the following commands:

```
> air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
> air_data$Temp_lag2<-NA
> air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
> View(air_data)
> m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
> m10 <-lm(Ozone~Ozone_lag1,data=air_data)
```

The environment pane on the right lists several objects: fit_res_ (List of 12), fit_tre_ (List of 12), fit_tre_ (List of 13), m1 (List of 13), and m10 (List of 13). The plot pane shows a time series plot of Ozone concentration from 1973.0 to 1973.4, with values ranging from 0 to 150. The plot shows a clear seasonal pattern with peaks around 1973.1 and 1973.3, and troughs around 1973.0 and 1973.4.



(Refer Slide Time: 05:12)



The screenshot shows an RStudio session with the following R code in the console:

```
air_data$Ozone_lag1<-NA
air_data$Ozone_lag2[3:n]<-air_data$Ozone[1:(n-2)]
View(air_data)
air_data$Temp_lag1<-NA
air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
View(air_data)
m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
m10 <-lm(Ozone~Ozone_lag1,data=air_data)
anova(m10,m1)
```

The ANOVA table output is as follows:

Analysis of Variance Table						
Model 1: Ozone ~ Ozone_lag1						
Model 2: Ozone ~ Ozone_lag1 + Temp_lag1						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	96	74010				
2	95	62352	1	11658	17.762	5.702e-05 ***

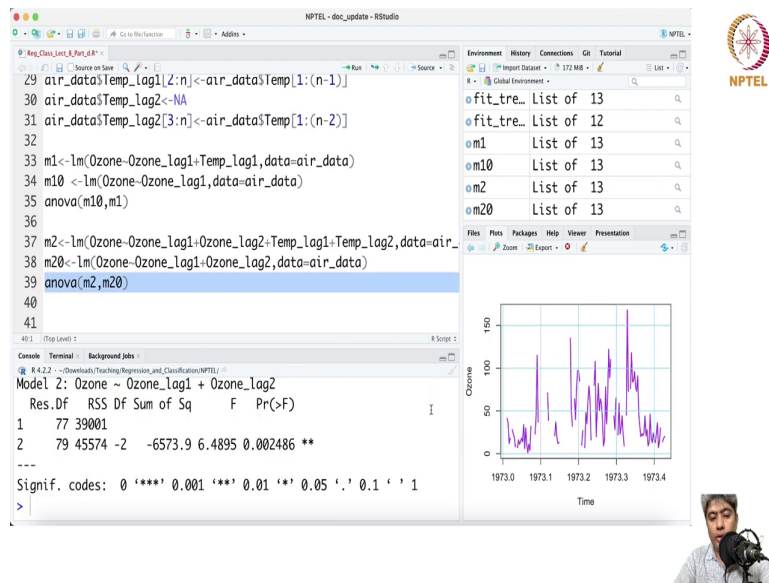
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The plot shows Ozone concentration over time from 1973.0 to 1973.4. The y-axis ranges from 0 to 150. The plot shows a highly volatile time series with several peaks reaching above 100.

Ozone_lag1 and Temperature_lag1 and similarly m1 say for null I will use 0, but just copy this entire thing and but instead of Temperature_lag1 I am dropping the Temperature_lag1. So, I am just saying the Ozone is only function of its own ok. Now, you see anova if you just run anova between m10 versus m1 ok.

You can see that it is the F test reject the null hypothesis and it says that the lag1 temperature does have effect on Ozone. So, you can do the Granger causal test in this way as well.

(Refer Slide Time: 06:30)



The screenshot shows the RStudio interface with the following R code in the script editor:

```
29 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
30 air_data$Temp_lag2<-NA
31 air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
32
33 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
34 m10 <-lm(Ozone~Ozone_lag1,data=air_data)
35 anova(m10,m1)
36
37 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_
38 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)
39 anova(m2,m20)
40
41
```

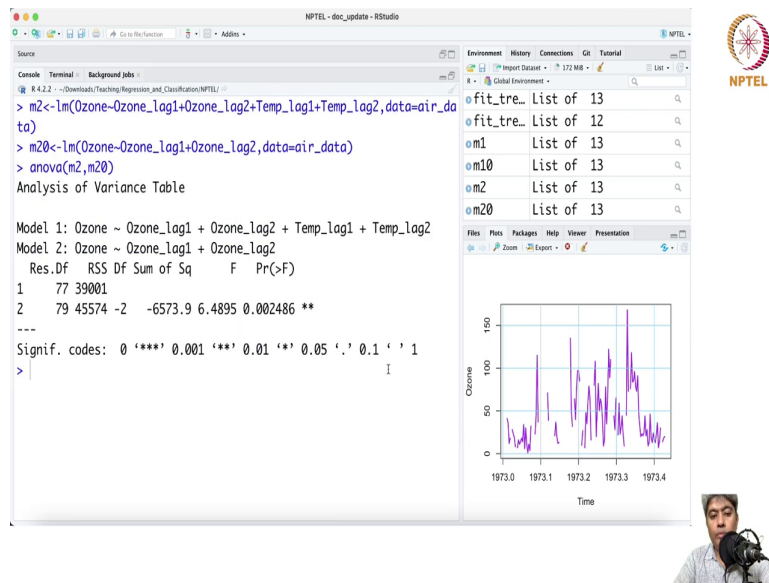
The console output shows the results of the ANOVA test for model m2:

```
Model 2: Ozone ~ Ozone_lag1 + Ozone_lag2
Res.Df  RSS Df Sum of Sq  F    Pr(>F)
1      77 39001
2      79 45574 -2   -6573.9  6.4895 0.002486 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The plot shows Ozone concentration over time from 1973.0 to 1973.4, with values ranging from 0 to 150. The plot is titled 'Ozone' and 'Time'.

Similarly, this is m2 this test with the 2 lag and then I can define the model with null model here, but only with the lag Ozone not with the temperature I am dropping the temperature and then we run the anova m test or F test 2 to comma m20 ok.

(Refer Slide Time: 07:11)





```
RStudio - NPTEL - doc_updates - RStudio
Source
Console Terminal Background Jobs
R 4.2.2 - (Downloads)Teaching/Regression_and_Classification/NPTEL/
> m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_da
ta)
> m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)
> anova(m2,m20)
Analysis of Variance Table

Model 1: Ozone ~ Ozone_lag1 + Ozone_lag2 + Temp_lag1 + Temp_lag2
Model 2: Ozone ~ Ozone_lag1 + Ozone_lag2
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
  1      77 39001
  2      79 45574 -2   -6573.9  6.4895 0.002486 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Environment History Connections Git Tutorial
Global Environment - 172 MB
fit_tre_ List of 13
fit_tre_ List of 12
m1 List of 13
m10 List of 13
m2 List of 13
m20 List of 13

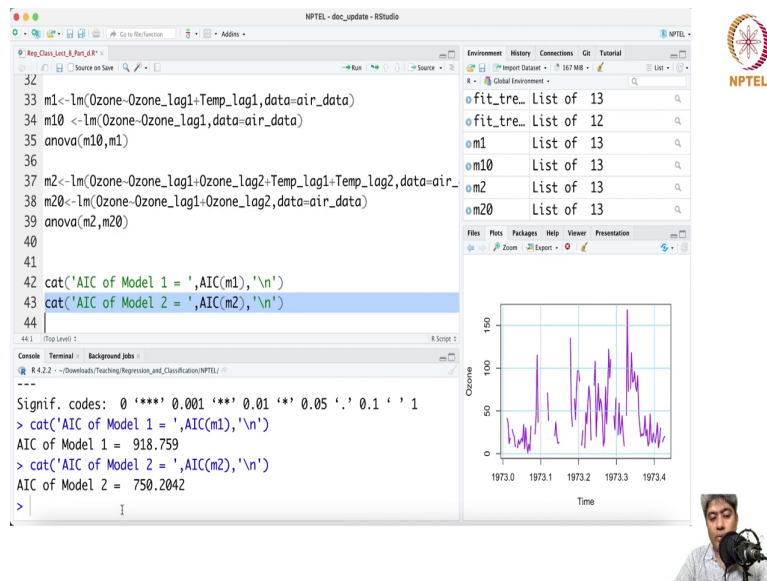
Ozone

Time



And you can see you can reject the F test. So, lag2 have effect on the Ozone.

(Refer Slide Time: 07:21)



The screenshot shows an RStudio window with the following R code in the script editor:

```
32 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)
34 m10 <-lm(Ozone~Ozone_lag1,data=air_data)
35 anova(m10,m1)
36
37 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_data)
38 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)
39 anova(m2,m20)
40
41
42 cat('AIC of Model 1 = ',AIC(m1),'\n')
43 cat('AIC of Model 2 = ',AIC(m2),'\n')
44
```

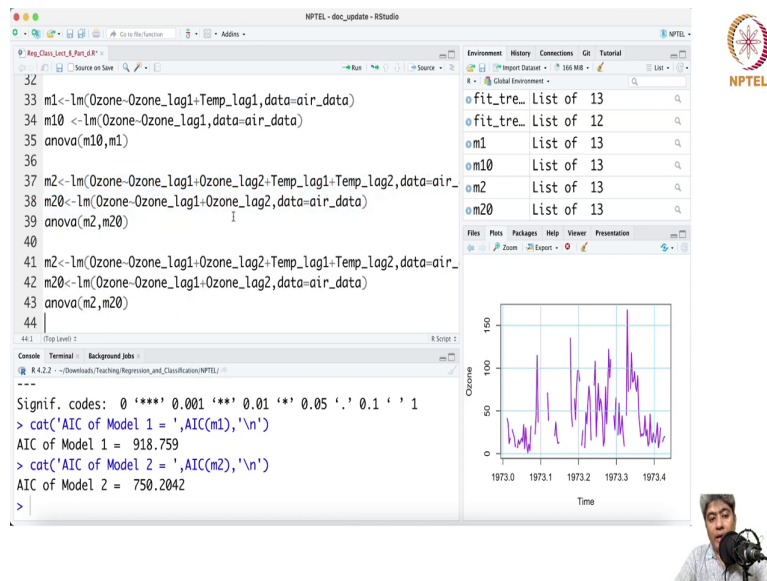
The console output shows the following results:

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> cat('AIC of Model 1 = ',AIC(m1),'\n')
AIC of Model 1 = 918.759
> cat('AIC of Model 2 = ',AIC(m2),'\n')
AIC of Model 2 = 750.2042
>
```

The environment pane on the right shows several objects: fit_tre_ (List of 13), fit_tre_ (List of 12), m1 (List of 13), m10 (List of 13), m2 (List of 13), and m20 (List of 13). The plot pane shows a time series plot of Ozone concentration over time, with the x-axis labeled 'Time' and the y-axis labeled 'Ozone'. The plot shows a highly volatile time series with values ranging from approximately 0 to 150. The NPTEL logo is visible in the top right corner of the RStudio window.

Now, if I compute the AIC of Model 1 and Model 2 clearly the Model 1 has a more AIC than the Model 2 the Model 2 has a lower mic AIC. So, we can we would prefer Model 2 over the Model 1.

(Refer Slide Time: 07:51)



The screenshot displays an RStudio interface with the following components:

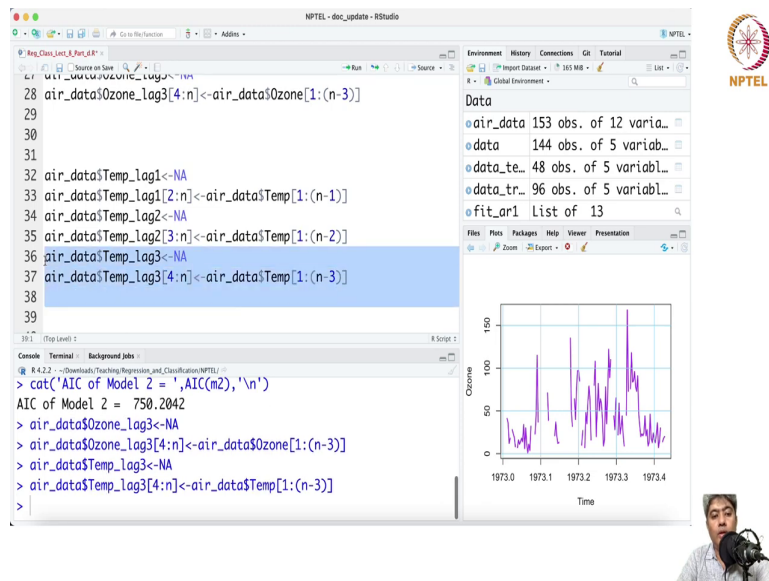
- Script Editor:** Contains R code for fitting two models and performing ANOVA tests:

```
32  
33 m1<-lm(Ozone~Ozone_lag1+Temp_lag1,data=air_data)  
34 m10 <-lm(Ozone~Ozone_lag1,data=air_data)  
35 anova(m10,m1)  
36  
37 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_...  
38 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)  
39 anova(m2,m20)  
40  
41 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_...  
42 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)  
43 anova(m2,m20)  
44
```
- Environment:** Lists several objects: fit_tre_ (List of 13), fit_tre_ (List of 12), m1 (List of 13), m10 (List of 13), m2 (List of 13), and m20 (List of 13).
- Console:** Shows the output of AIC calculations:

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
> cat('AIC of Model 1 = ',AIC(m1),'\n')  
AIC of Model 1 = 918.759  
> cat('AIC of Model 2 = ',AIC(m2),'\n')  
AIC of Model 2 = 750.2042  
>
```
- Plots:** A time-series plot of Ozone levels from 1973.0 to 1973.4, showing a fluctuating purple line with values between 0 and 150.

Similarly, you can go for the third variable third lag we have to create a lag here. So, we can always create these lines.

(Refer Slide Time: 08:00)



The screenshot shows the RStudio interface. The script editor contains the following R code:

```
28 air_data$Ozone_lag3[4:n]<-air_data$Ozone[1:(n-3)]
29
30
31
32 air_data$Temp_lag1<-NA
33 air_data$Temp_lag1[2:n]<-air_data$Temp[1:(n-1)]
34 air_data$Temp_lag2<-NA
35 air_data$Temp_lag2[3:n]<-air_data$Temp[1:(n-2)]
36 air_data$Temp_lag3<-NA
37 air_data$Temp_lag3[4:n]<-air_data$Temp[1:(n-3)]
38
39
```


The console shows the following output:

```
> cat('AIC of Model 2 = ',AIC(m2),'\n')
AIC of Model 2 = 750.2042
> air_data$Ozone_lag3<-NA
> air_data$Ozone_lag3[4:n]<-air_data$Ozone[1:(n-3)]
> air_data$Temp_lag3<-NA
> air_data$Temp_lag3[4:n]<-air_data$Temp[1:(n-3)]
>
```

The Data pane on the right shows the following data objects:

- air_data: 153 obs. of 12 variables
- data: 144 obs. of 5 variables
- data_te: 48 obs. of 5 variables
- data_tr: 96 obs. of 5 variables
- fit_ar1: List of 13

The Plots pane shows a time series plot of Ozone vs Time, with the y-axis ranging from 0 to 150 and the x-axis ranging from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks reaching above 100.



First, we have to create these lines. So, maybe lag3, lag3, 2, 4, 3.

(Refer Slide Time: 08:26)

The screenshot shows an RStudio interface with the following components:

- Source Editor:** Contains R code for creating lags of ozone and temperature data, fitting models, and performing ANOVA tests.


```

44
45 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_
46 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)
47 anova(m2,m20)
48
49 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag
50 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)
51 anova(m2,m20)
52
53 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag
54 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)
55 anova(m2,m20)
56

```
- Environment:** Lists objects created in the environment, including `fit_tre`, `m1`, `m10`, `m2`, `m20`, and `m3`, each being a "List of 13".
- Console:** Shows the execution of the code in the source editor.


```

R 4.2.2 -> (Downloads)\Teaching\Regression_and_Classification\NPTEL\
> air_data$Ozone_lag3<-NA
> air_data$Ozone_lag3[4:n]<-air_data$Ozone[1:(n-3)]
> air_data$Temp_lag3<-NA
> air_data$Temp_lag3[4:n]<-air_data$Temp[1:(n-3)]
> m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag2+
Temp_lag3,data=air_data)
>

```
- Plots:** An environmental plot showing "Ozone" concentration on the y-axis (ranging from 0 to 150) against "Time" on the x-axis (ranging from 1973.0 to 1973.4). The plot displays a highly volatile time series.

So, we have to create these lags for ozone and then now we have to create these lags for temperature third lag plus 3. And then what we need is model for third lag and Ozone plus 3 lag plus 3 and ok.

(Refer Slide Time: 09:10)

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code for fitting linear models and performing ANOVA tests. The code includes:

```
48  
49 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_  
50 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)  
51 anova(m2,m20)  
52  
53 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag  
54  
55 m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3,data=air_data)  
56 anova(m3,m30)  
57  
58  
59
```
- Environment:** Lists several objects: m1, m10, m2, m20, m3, m30, each associated with a 'List of 13'.
- Console:** Shows the output for 'Model 2: Ozone ~ Ozone_lag1 + Ozone_lag2 + Ozone_lag3'. The output is as follows:

```
Model 2: Ozone ~ Ozone_lag1 + Ozone_lag2 + Ozone_lag3  
Res.Df  RSS Df Sum of Sq  F Pr(>F)  
1      61 26091  
2      64 27174  -3    -1083.3  0.8443  0.475
```
- Plots:** A time-series plot of Ozone levels from 1973.0 to 1973.4. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time'.

And the 3 naught null we just Ozone it will be only function of Ozone and then we will see if it is still effective.

(Refer Slide Time: 09:34)

```
Source
R 4.2.2 -> (Downloads)TeachingRegression_and_Classification(NPTEL)
> cat("AIC of Model 1 = ",AIC(m1),'\n')
AIC of Model 1 = 918.709
> cat("AIC of Model 2 = ",AIC(m2),'\n')
AIC of Model 2 = 750.2042
> air_data$Ozone_lag3<-NA
> air_data$Ozone_lag3[4:n]<-air_data$Ozone[1:(n-3)]
> air_data$Temp_lag3<-NA
> air_data$Temp_lag3[4:n]<-air_data$Temp[1:(n-3)]
> m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag2+
Temp_lag3,data=air_data)
> m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3,data=air_data)
> anova(m3,m30)
Analysis of Variance Table

Model 1: Ozone ~ Ozone_lag1 + Ozone_lag2 + Ozone_lag3 + Temp_lag1 +
Temp_lag2 +
Temp_lag3
Model 2: Ozone ~ Ozone_lag1 + Ozone_lag2 + Ozone_lag3
Res.Df  RSS Df Sum of Sq    F Pr(>F)
1      61 26091
2      64 27174  -3   -1083.3  0.8443  0.475
>
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	61	26091			
2	64	27174	-3	-1083.3	0.8443 0.475

And now when we are doing third lag we are adding third lag its not effective anymore.

(Refer Slide Time: 09:38)

The screenshot shows an RStudio session with the following code in the script editor:

```
51 anova(m2,m20)
52
53 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag
54
55 m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3,data=air_data)
56 anova(m3,m30)
57
58
59
60 cat('AIC of Model 1 = ',AIC(m1),'\n')
61 cat('AIC of Model 2 = ',AIC(m2),'\n') ;
62 cat('AIC of Model 3 = ',AIC(m3),'\n')
63
```

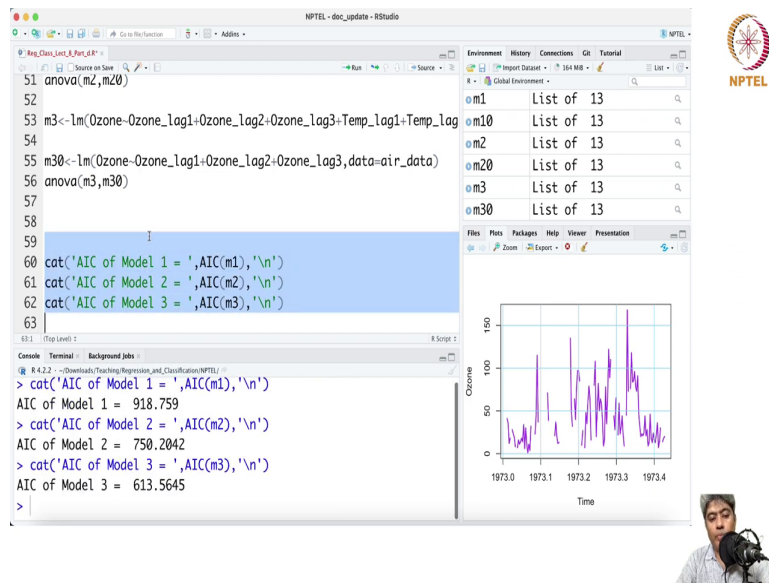
The console output shows the following results for Model 2:

```
Model 2: Ozone ~ Ozone_lag1 + Ozone_lag2 + Ozone_lag3
Res.Df  RSS Df Sum of Sq  F Pr(>F)
1      61 26091
2      64 27174 -3    -1083.3  0.8443  0.475
> cat('AIC of Model 3 = ',AIC(m3),'\n')
AIC of Model 3 = 613.5645
>
```

The environment pane on the right shows several objects: m1, m10, m2, m20, m3, and m30, each with a 'List of 13' items. The plot pane shows a time series plot of Ozone concentration from 1973.0 to 1973.4, with values ranging from 0 to 150.

But can we choose m3 what is the AIC of model third model.

(Refer Slide Time: 09:56)



The screenshot displays the RStudio interface with the following code in the script editor:

```
51 anova(m2,m20)
52
53 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag
54
55 m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3,data=air_data)
56 anova(m3,m30)
57
58
59
60 cat('AIC of Model 1 = ',AIC(m1),'\n')
61 cat('AIC of Model 2 = ',AIC(m2),'\n')
62 cat('AIC of Model 3 = ',AIC(m3),'\n')
63
```

The console output shows the following results:

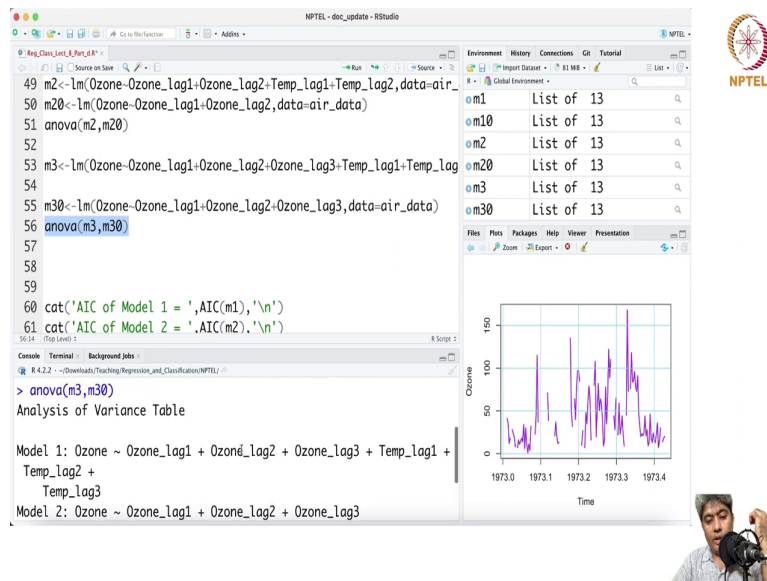
```
> cat('AIC of Model 1 = ',AIC(m1),'\n')
AIC of Model 1 = 918.759
> cat('AIC of Model 2 = ',AIC(m2),'\n')
AIC of Model 2 = 750.2042
> cat('AIC of Model 3 = ',AIC(m3),'\n')
AIC of Model 3 = 613.5645
>
```

The environment pane on the right lists several objects: m1, m10, m2, m20, m3, and m30, each associated with a 'List of 13'.

A line plot titled 'Ozone' vs 'Time' is shown in the bottom right. The x-axis represents 'Time' from 1973.0 to 1973.4, and the y-axis represents 'Ozone' from 0 to 150. The plot shows a highly volatile time series with multiple peaks and troughs.

Now, if you interestingly if you see what is happening. You see what is happening here the AIC is constantly dropping whereas, model; that means, third model is probably better than the first and second model.

(Refer Slide Time: 10:21)



The screenshot shows an RStudio window with the following R code in the script editor:

```
49 m2<-lm(Ozone~Ozone_lag1+Ozone_lag2+Temp_lag1+Temp_lag2,data=air_data)
50 m20<-lm(Ozone~Ozone_lag1+Ozone_lag2,data=air_data)
51 anova(m2,m20)
52
53 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag2+Temp_lag3,data=air_data)
54
55 m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3,data=air_data)
56 anova(m3,m30)
57
58
59
60 cat('AIC of Model 1 = ',AIC(m1),'\n')
61 cat('AIC of Model 2 = ',AIC(m2),'\n')
```

The console output shows the results of the ANOVA test for Model 3:

```
> anova(m3,m30)
Analysis of Variance Table

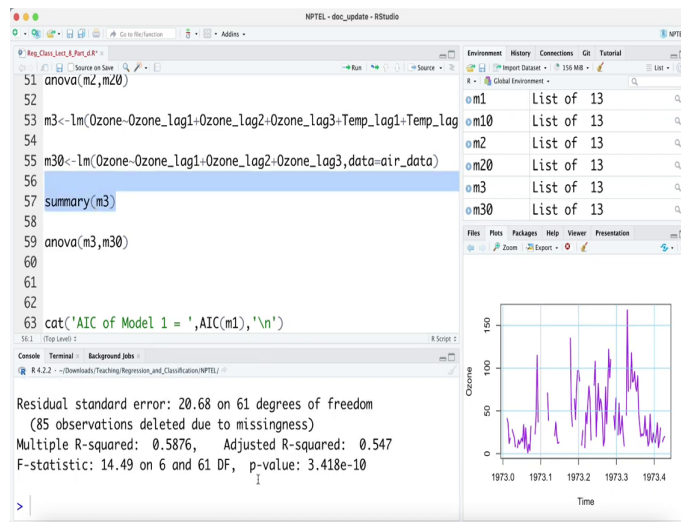
Model 1: Ozone ~ Ozone_lag1 + Ozone_lag2 + Ozone_lag3 + Temp_lag1 +
  Temp_lag2 +
  Temp_lag3
Model 2: Ozone ~ Ozone_lag1 + Ozone_lag2 + Ozone_lag3
```

The plot on the right shows Ozone concentration over time from 1973.0 to 1973.4. The y-axis ranges from 0 to 150. The plot shows a highly volatile time series with several peaks reaching above 100.

However, what we are seeing that third model the lag does not have any effect. So, that means, we cannot really use we cannot say that third lag has any effect on the 3 lag model of a granger causal model does any effect any of temperature has no effect on the Ozone.

So, that means, as we are putting as we are you know put more and more lags. So, naturally what is happening the model complexity increases and its going to higher and higher dimension as model complexity increases it is doing some overfitting because my AIC is constantly going down. But we know, but most likely it is effectively losing its interpretability and that is why probably its not running its not giving any effect.

(Refer Slide Time: 11:21)



The screenshot displays the RStudio interface with the following content:

```
51 anova(m2,m20)
52
53 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag
54
55 m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3,data=air_data)
56
57 summary(m3)
58
59 anova(m3,m30)
60
61
62
63 cat('AIC of Model 1 = ',AIC(m1),'\n')
```

The console output shows the following statistical results:

```
Residual standard error: 20.68 on 61 degrees of freedom
(85 observations deleted due to missingness)
Multiple R-squared: 0.5876, Adjusted R-squared: 0.547
F-statistic: 14.49 on 6 and 61 DF, p-value: 3.418e-10
```

The environment pane on the right lists several objects: m1, m10, m2, m20, m3, and m30, each associated with a 'List of 13'.

The plot pane shows a time series plot of Ozone levels over time. The y-axis is labeled 'Ozone' and ranges from 0 to 150. The x-axis is labeled 'Time' and ranges from 1973.0 to 1973.4. The plot shows a highly volatile time series with several peaks reaching above 100.



(Refer Slide Time: 11:32)

The screenshot shows an RStudio session with the following code in the script editor:

```
51 anova(m2, m20)
52
53 m3<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3+Temp_lag1+Temp_lag2)
54
55 m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3, data=air_data)
56
57 summary(m3)
58
59 anova(m3, m30)
60
61
62
63 cat('AIC of Model 1 = ', AIC(m1), '\n')
```

The console output shows the following table:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-57.075641	39.207726	-1.456	0.15060
Ozone_lag1	0.416355	0.137841	3.021	0.00368 **
Ozone_lag2	-0.008118	0.143069	-0.057	0.95493
Ozone_lag3	0.187251	0.142906	1.310	0.19500
Temp_lag1	0.294474	0.584445	0.504	0.61618
Temp_lag2	0.244689	0.577099	0.424	0.67306

The plot on the right shows Ozone levels over time (1973.0 to 1973.4), with values ranging from 0 to 150. The plot is a time-series plot with a purple line representing the Ozone concentration over time.



But we thought we know that you know lag1 and lag2 does have a effect. So, we can see that you know you know very high M3 in case of third model except the lag1 model does not have any effect.

(Refer Slide Time: 11:49)

The screenshot displays the RStudio interface. The script editor contains the following R code:

```
54  
55 m30<-lm(Ozone~Ozone_lag1+Ozone_lag2+Ozone_lag3,data=air_data)  
56  
57 summary(m3)  
58  
59 summary(m2)  
60 anova(m3,m30)  
61  
62  
63  
64 cat('AIC of Model 1 = ',AIC(m1),'\n')  
65 cat('AIC of Model 2 = ',AIC(m2),'\n')  
66 cat('AIC of Model 3 = ',AIC(m3),'\n')
```

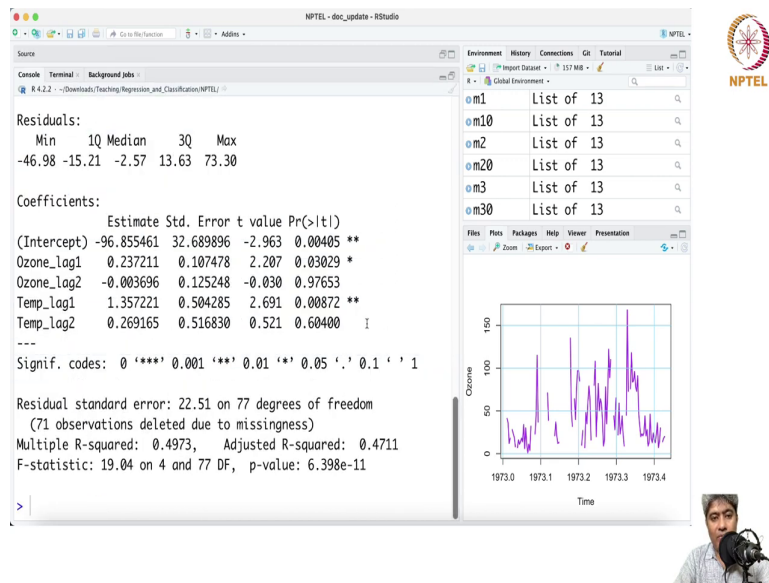
The console output shows the following statistical results:

```
Residual standard error: 22.51 on 77 degrees of freedom  
(71 observations deleted due to missingness)  
Multiple R-squared: 0.4973, Adjusted R-squared: 0.4711  
F-statistic: 19.04 on 4 and 77 DF, p-value: 6.398e-11
```

The environment pane on the right lists several objects: m1, m10, m2, m20, m3, and m30, each associated with a 'List of 13' items. The plot pane shows a time-series plot of 'Ozone' levels from 1973.0 to 1973.4, with values ranging from 0 to 150. The plot shows a highly volatile time series with a clear upward trend.

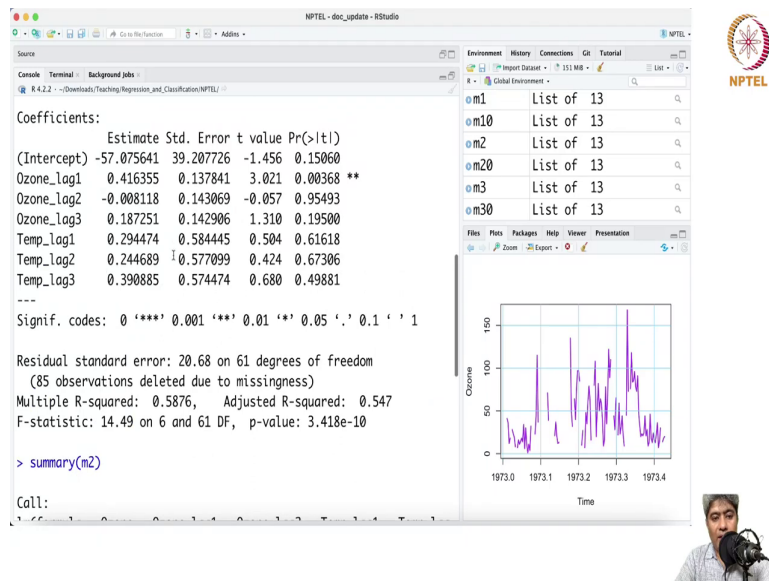


(Refer Slide Time: 11:53)



Whereas, m2 if you see m2 lag1 temperature does have a effect we see it does have a effect whereas, the lag2 does not have a effect.

(Refer Slide Time: 12:00)



Also, if you look into the standard error standard error in third model is 0.577 whereas, in the second model for lag 1 is 0.50. So, that means, standard error is increasing. So, there is a high possibility that a multicollinearity also creeping in because of the multicollinearity these lags are not any more effective because their standard error is going up because of the multicollinearity.

So, you have to be very careful about how you do the interpretation of these statistical inference and these you have to be very careful about these statistical inferences when particularly when you are adding more and more features just because every time you are increasing lag.

That means, your model is getting complex your complexity of your model is increasing it will have a higher tendency to overfit because end of the day your training datasets is finite

most of the time your data training dataset is not increasing. So, as a result your model will overfit and as it overfits you do not want your model to overfit because your bias will when it happens your bias will be small.

But you will have a very high variance in out of the sample it will not do very well. So, you have to be very careful about model fitting when for adding more complex lag variable and it is better to obtain a parsimonious small model and just stop there, ok. So, far this week we will this is that is how this much we will be discussing. Next week we will see you with a new video with a new chapter.

Thank you very much. Take care. Bye.