

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

Lecture - 28
Granger Casual model

Welcome to the part c of lecture series 8. In this lecture, we are going to talk about Granger Causal Inference.

(Refer Slide Time: 00:26)



Correlation and Causation

- ▶ "Correlation does not imply causation"
- ▶ Why causation is important with respect to predictive analytics?
- ▶ Suppose we are modelling

$y = f(x_1, x_2)$

If we know x_1 or x_2 has causal effect on y , then we will be confident about the predictive power of the model.

- ▶ However, if x_1 or x_2 does not have a causal effect on y , and what we observe a spurious correlation, then the model



In this lecture, first we will try to understand that, "Correlation does not imply causation". This is a very important statement and question is always why causation is important with respect to predictive analytics.

So, suppose, we are modelling y equal to some function of x_1 and x_2 . So, x_1 and x_2 are two features, x_1 and x_2 are features and you want to predict the value of y . If we know x_1 or x_2 has a causal effect of y , then we will be confident about the predictive power of the model. However, if the x_1 and x_2 does not have a causal effect of on y and what we observe a spurious correlation, a spurious correlation, then the model will fail in live production environment.

So, and that is the time what happens is model do not generalize it. So, you may heard if you are coming from machine learning background or engineering background, that model do not generalizes in the live production environment.

Why it so? Because if you find a predictor which may be correlated in some for some spurious reason. And remember that, what is correlation? Correlation is just a formula. End of the day, correlation is just a formula which for whatever reason is showing, maybe it may show a high correlation between two variable.

But if you know there is a physical phenomenon, there is a there is a domain relation, there is the domain do explain the relationship between a feature and the response. Then, you know whether it is a linear, non-linear whatever be the relationship, you from the feature you can explain or you can expect why to behave in certain way. Maybe with some uncertainty, but you know the there is a relationship. In that case, your reliance on your predictive model increases.

(Refer Slide Time: 03:27)

The slide is titled "Regression Model for Granger Causality". It features a list of three bullet points, each enclosed in a red hand-drawn box. The first bullet point is "In practice, it is difficult to answer causal questions." and has "Randomized case-control experiment" and "RCT" written above it. The second bullet point is "Granger causality can be used to make causal statements." The third bullet point is "Naturally, Granger causality helps us to understand if one time series is useful for predicting another". Below the list is a "Question" box containing the text "Does one time series cause another, controlling for lags?". The NPTEL logo is in the top right corner, and a small video inset of a speaker is in the bottom right corner.

Regression Model for Granger Causality

Randomized case-control experiment
RCT

- ▶ In practice, it is difficult to answer causal questions.
- ▶ Granger causality can be used to make causal statements.
- ▶ Naturally, Granger causality helps us to understand if one time series is useful for predicting another

Question: Does one time series cause another, controlling for lags?

So, how can we; in practice it is difficult to answer the causal question. Why? Because it is effectively very difficult. Often time what we see is just a observed response. Proving causality is very difficult proposition. One possible a possibility is randomize experiment, randomize, randomized case control, case control experiment.

Often time, in economics nowadays its very popular RCT, randomized case treatment studies or causal treatment studies. It is well known in the statistics literature for about 100 years now. And in, this randomized case control experiment might help us establishing the causal relationship between a feature and a response. But there could be a situation where simple ethical construct, ethical; for ethical reason you cannot run a randomized case control experiment.

For example, if you want to; how a randomized case control experiment happens? So, for example, you want to check the efficacy of a vaccine. So, what you will do? Randomly, you will some patient will come in the study. You will toss a coin, if it is head a patient will get a vaccine. If the tail, then the patient will get a placebo. And you do it for say few 100 patients.

And at the end of the day, you will observe the patients for a year or 2 year or 3 years. And then, you will see that how many of the case group have developed the disease and how many of the control group has developed the disease. So, and that will give you a sense of the whether the vaccine is effective or not.

And, but the same study you cannot run on certain experimental situation. For example, you want to test if smoking causes cancer. Now, you have recent, we have enough evidence then very likely smoking do causes cancer.

Now, you cannot just toss a coin you, say that ok you group of people will do smoke and you cannot smoke. If you know what treatment may cause harm to a group of people, then you cannot run that study. So, randomized case control experiment has its limitation.

Now, Granger causality can be used to make some causal treatment with some with some limited effect. So, Granger causality helps us to understand if one time series is useful for predicting another. So, question that Granger causality try to answer, does one time series cause another controlling for lags. This is the question that we try to answer, we try to figure out.



(Refer Slide Time: 07:30)

Regression Model for Granger Causality

- ▶ Basic univariate Granger causality test:
- ▶ We have two time series $\{(y_t, x_t) | t = 1, 2, \dots, n\}$
- ▶ **Question:** Are lags of x_t predictive of y_t controlling for lags of y_t ?

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_k y_{t-k} + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \dots + \gamma_k x_{t-k} + \epsilon_t,$$

where we assume $\mathbb{E}(\epsilon_t | \mathcal{F}_{t-1}) = 0$



So, basic we start with the basic univariate Granger causality test. We have two time series a y_t and x_t , and t is from say 1, 2 up to n . So, t is the time point. Are the lags of x_t predictive of y_t controlling for lags of y_t ? Are the lags of x_t can predict the y_t for the lags controlling the lags of y_t ?

So, how you do that? You create y_t , you want to predict y_t equal to β_0 , $\beta_1 y_{t-1}$, $\beta_2 y_{t-2}$ plus dot dot dot $\beta_k y_{t-k}$, plus $\gamma_1 x_{t-1}$ plus $\gamma_2 x_{t-2}$ plus $\gamma_k x_{t-k}$. You see this model is completely linear in parameter. We can easily find the estimates using OLS estimator or any standard methodology, like Lasso or elastic net.

Now, what this model means? This model means basically that after taking. So, whatever the variability in y_t , many of the variability getting explained by its own historical behaviour of y_{t-1} , y_{t-2} up to y_{t-k} .

After that, after that if still there are some residuals that can be explained by x_{t-1} or x_{t-2} or x_{t-k} . Then, we can say that a lag variable of x_{t-k} or some lag variable of x time series do have some effect on y or we can say some lag variable of x_{t-k} do have some effect on y_t .

(Refer Slide Time: 09:31)

▶ **Question:** Are lags of x predictive of y , controlling for lags of y ?

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_k y_{t-k} + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \dots + \gamma_k x_{t-k} + \epsilon_t$$

where we assume $\mathbb{E}(\epsilon_t | \mathcal{F}_{t-1}) = 0$

$x \longrightarrow y$
 $x(t-k) \longrightarrow y(t)$


Regression Model for Granger Causality

▶ Here \mathcal{F}_{t-1} summarizes the information up to time $(t-1)$ of



So, this is our, we can say that Granger causes y_t .

(Refer Slide Time: 09:48)

Regression Model for Granger Causality




- ▶ Here \mathcal{F}_{t-1} summarizes the information up to time $(t-1)$ of both x and y
- ▶ H_0 : $\gamma_1 = \gamma_2 = \dots = \gamma_k = 0$
- vs
- ▶ H_a : $\gamma_i \neq 0$ at least one lag of x provides additional information.
- ▶ We run the F-test



Now, how we do that; if so, what we do a test for Granger causal Granger causality. If the gamma 1, gamma 2, gamma k, what are the gamma? Gamma remember that gamma is these coefficient, these coefficients.


These coefficients are all 0; that means, none of the x have any effect on the y . So, y is only function of its own historical behaviour. And if any one of the gamma e at least one lag of x is additional provide some additional information to y , so any at least one gamma is non-zero. That means, at least one lag any one lag of x do have effect on y t. So, if that happens then we can say that time series x Granger causes time series y .

(Refer Slide Time: 10:48)



How do we choose the number of lags?

- ▶ It is a tradeoff of between the bias vs statistical power.
- ▶ With too few lags, we can find residual autocorrelation. It may gives us a biased test.
- ▶ With too many lags, we might incorrectly reject the null due to spurious correlation.



Now, how do we choose the number of lags? It is a tradeoff between bias and versus the statistical power. With too few lags, we can find residuals autocorrelation. It may give us a bias test. With too many lags, it might correctly incorrectly reject the null due to just some spurious correlation.

So, again, what you have to do? You have to choose different for values of k , you have to do the test and also you have to calculate the mean square error. And you have to check in the out of the sample whether what is the RMSc or what is the adjusted R square of the data. And then, you have to choose a correct kind of lag where you should stop. And for that lag model, if the x time series has any effect on the y time series.




(Refer Slide Time: 11:51)

Is it Causality?

From the statistical test, can we conclude that the x causes the future number of y ? **There are several potential issues when making causal statements:**

- ▶ **Confounders:** There may be some other variable z , which is correlated with x , and that is the true cause of y .
- ▶ **Lead-lag relationship / feedback loop**

```
graph LR; X1((X_{t-1})) --> Y1((Y_t)); Y1 --> X2((X_{t+1})); Y2((Y_{t-1})) --> X3((X_t)); X3 --> Y3((Y_{t+1}));
```
- ▶ **Spurious Correlation:** A correlation between the two variables, but it is coincidental !!

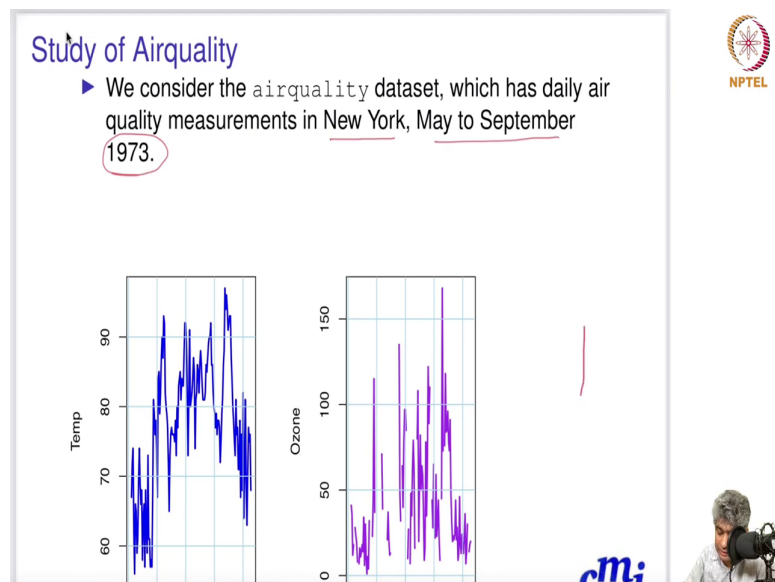
But the question is ok, this is just a test statistical test we are talking about. But from the statistical test, can we conclude that x causes the future number of y or x causes y .

There are several potential issue when making causal statement. First issue could be confounder. There may be some other variable z , which is correlated with x , and that is the true cause of y . And what you are seeing is just all the you know some spurious correlation. It is not really the mean correlation. There could be a lead drag relationship or feedback loop, like x_{t-1} cause y_t and y_t causes x_{t+1} .

Similarly, y_{t-1} cause x_t and x_t cause y_{t+1} . And this feedback loop would be a problem too. Then, who causing whom, is it or the causality is both way, you have to conclude in that way. So, and the final is spurious correlation between the two variable, but it

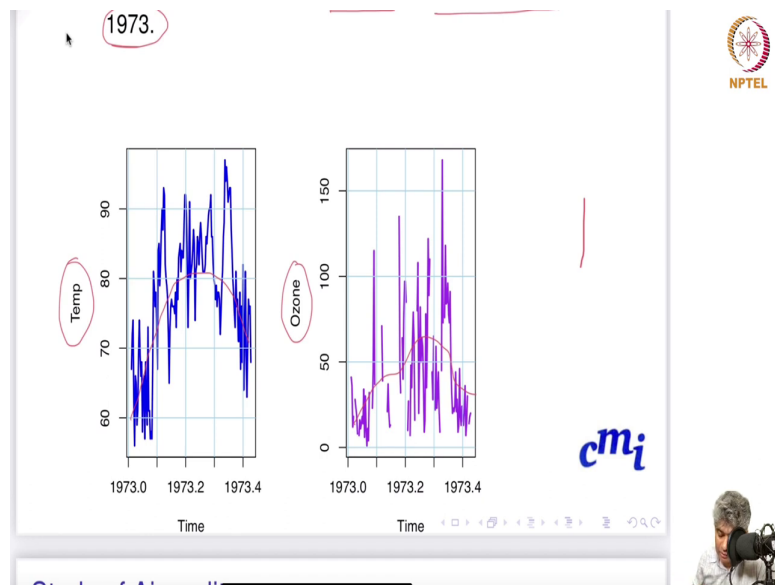
is just it was just a coincidental. So, you have to be very careful when you are making such causes.

(Refer Slide Time: 13:17)



So, there is a airquality dataset, which has daily airquality measurement in New York from May to September 1973.

(Refer Slide Time: 13:32)





Here is the dataset, the and the ozone. In that there are two variable, one is ozone variable another is the temperature, ok. How the temperature behaves and during that time how the you know ozone actually behave.

(Refer Slide Time: 13:51)

Study of Airquality

```
> library(lmtest)
> cat('Model 1', '\n')
Model 1
> grangertest(Ozone ~ Temp, order = 1, data = airqu
Granger causality test

Model 1: Ozone ~ Lags(Ozone, 1:1) + Lags(Temp, 1:1)
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
```



So, question is if temperature causes the movement of the ozone, it does have the ozone level, and ozone level has some, ozone is a greenhouse gas. So, question is whether can be is temperature going to do causing the level of ozone.

So, we there in library lm test, there is a function called Granger test. You we can use that. Just simply ozone follow temperature, order 1, data equal to airquality and then it simply runs the model 1 ozone and then you can see the P value for the test is extremely small.

So, we reject the null hypothesis that none of them have any effects. So, that means, the temperature does not have a effect. So, we reject that.




(Refer Slide Time: 14:53)

```
library(lmtest) ✓
cat('Model 1', '\n')
Model 1
> grangertest(Ozone ~ Temp, order = 1, data = airqu)
Granger causality test

Model 1: Ozone ~ Lags(Ozone, 1:1) + Lags(Temp, 1:1)
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
```

~~H₀: Temp does not have an effect on Ozone~~
vs
H_a: Temp does have effect on Ozone

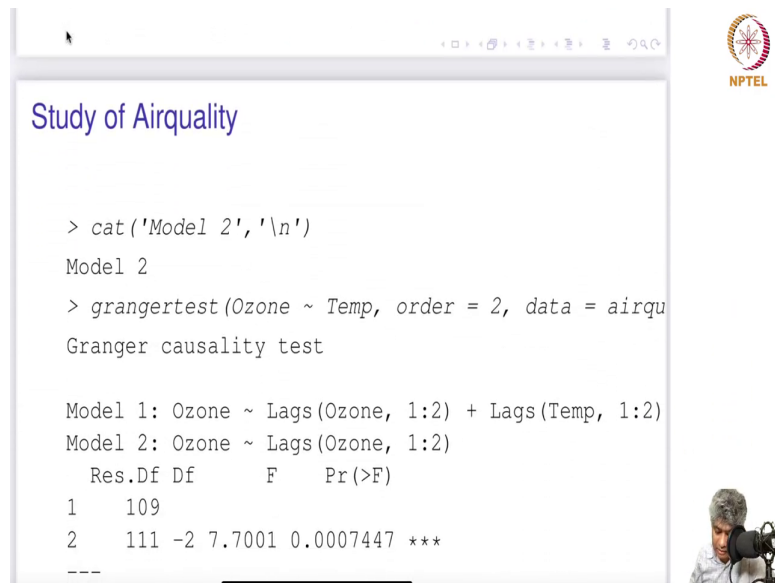
$p < 0.001$
we reject H₀



So, what is null? Null is temperature, does not have a effect, does not have an effect on ozone versus alternative temperature. Does have effect on ozone.

So, since P value is too small, P value is too small, so we reject we reject null hypothesis. That is we reject this guy, and we say that temperature does have a effect on the ozone.

(Refer Slide Time: 15:44)




Study of Airquality

```
> cat('Model 2', '\n')
Model 2
> grangertest(Ozone ~ Temp, order = 2, data = airqu)
Granger causality test

Model 1: Ozone ~ Lags(Ozone, 1:2) + Lags(Temp, 1:2)
Model 2: Ozone ~ Lags(Ozone, 1:2)
  Res.Df Df    F   Pr(>F)
1     109
2     111 -2 7.7001 0.0007447 ***
---
```

NPTEL



Then, it was this model was done with only one lag. What about going for lag 2? And we did the; we did it with the lag 2 here, and now still it rejects the null hypothesis.

(Refer Slide Time: 16:04)

Study of Airquality

✓ AIC of Model 1 = 918.759

✓ AIC of Model 2 = 750.2042 ✓

NPTEL

And then we calculated the AIC of model 1 and AIC of model 2. Between the model 1 and model 2, this is better model because AIC is minimum this case. So, we should choose the our inference from the model 2.

(Refer Slide Time: 16:24)

Next week ...

- We will do some hands-on...

NPTEL

cmi

So, now we will stop here, and we will go to the do some hands on. So, see you in the next video with hands on.

Thanks.