Predictive Analytics
Regression and Classification

Lecture 5 : Part 2

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2020

Correlation and Causation

- "**Correlation does not imply causation**"
- Why causation is important with respect to predictive analytics?
- Supppose 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 will fail in the live production environment.

Regression Model for Granger Causality

- 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?

Regression Model for Granger Causality

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

$$\begin{aligned} y_t = \beta_0 &+ \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} \\ &+ \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \cdots + \gamma_k x_{t-k} + \epsilon_t, \end{aligned}$$

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

Regression Model for Granger Causality

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

- $H_0 : \quad \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0$

vs

- $H_a : \quad \gamma_i \neq 0$ at least one lag of x provides additional information.

- We run the F-test

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.

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**:

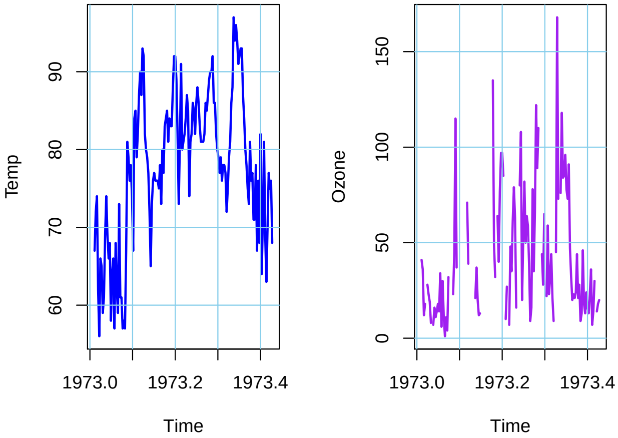
- **Confounders**: There may be some other variable z, which is correlated with x,and that is the true cause of y.
- **Lead-lag relationship / feedback loop**.

$$\begin{aligned} x_{t-1} \rightarrow \quad & y_t \quad \rightarrow x_{t+1} \\ y_{t-1} \rightarrow \quad & x_t \quad \rightarrow y_{t+1} \end{aligned}$$

- **Spurious Correlation**: A correlation between the two variables, but it is coincidental !!

Study of Airquality

- We consider the `airquality` dataset, which has daily air quality measurements in New York, May to September 1973.



Study of Airquality

```
> library(lmtest)
> cat('Model 1','\n')

Model 1

> grangertest(Ozone ~ Temp, order = 1, data = airquality)

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
```

Study of Airquality

```
> cat('Model 2','\n')
Model 2
> grangertest(Ozone ~ Temp, order = 2, data = airquality)
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 ***
---
Signif. codes:  0
```

Study of Airquality

```
AIC of Model 1 =  918.759
AIC of Model 2 =  750.2042
```

Next week ...

- We will so some hands-on...

Thank You

sourish@cmi.ac.in