Non i.i.d. data

Wenda Zhou

June 20, 2017

Non i.i.d. data

The i.i.d. assumption

Most statistical methods assume that the observations are independent and identically distributed (conditional on the covariates and parameters).

Non i.i.d. data

The i.i.d. assumption

Most statistical methods assume that the observations are independent and identically distributed (conditional on the covariates and parameters).

Non i.i.d. data

Also common to encounter data that violates this assumption.

- Time series (e.g. financial or economic data)
- Cross-sectional data.
- Panel and longitidunal data (e.g. cohort studies)

Time series

Time series data arise when the present depends on the past.

Stationarity

In order for our statistical analysis to make sense, the future must look like the past. This condition is called stationarity.

Integrated time series

Sometimes the case that a time series is not stationary, but its difference series is. We say the time series is integrated.

Autocorellation and partial autocorrelation

In a time series, natural to measure correlation with "self" in past.

Autocorellation function (acf)

$$acf(i) = \text{correlation between now and } i \text{ days in past} \tag{1}$$

Partial autocorrelation function (pacf)

By stationarity, if have some correlation at period 1, will also have correlation at period 2.
The *pacf* measures the additional correlation compared to that expected.

Auto-regressive models

Most common time series model.
An autoregressive model of order p is defined as:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \epsilon_t, \tag{2}$$

where $\alpha_1, \ldots, \alpha_p$ are the parameters of the model, and ϵ_t is a random noise ("innovation").

Auto-regressive models

Most common time series model.
An autoregressive model of order p is defined as:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \epsilon_t, \tag{2}$$

where $\alpha_1, \ldots, \alpha_p$ are the parameters of the model, and ϵ_t is a random noise ("innovation").
Can be viewed as linear regression of the present on p past time points.

Moving-average models

Also commonly used time series model.
A moving average model of order q is defined as:

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{3}$$

Moving-average models

Also commonly used time series model.
A moving average model of order q is defined as:

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{3}$$

Can be viewed as linear regression on the past q innovations (that are not observed).

ARIMA

General time series modelling strategy - auto-regressive integrated moving average.
Most general commonly used model – good choice to do predictions on time series.
Need to supply (or select) three parameters:

- p Order of the auto-regressive component
- d Order of the differencing
- q Order of the moving-average component

ARIMA

General time series modelling strategy - auto-regressive integrated moving average.
Most general commonly used model – good choice to do predictions on time series.
Need to supply (or select) three parameters:

- p Order of the auto-regressive component
- d Order of the differencing
- q Order of the moving-average component

Formulas get pretty ridiculous . . .

Seasonality and exogenous variables

Seasonality

Time series often display some cyclical behaviour – seasonality. Important aspect to model – often know the length of the seasonality.

Exogenous variables

Some time series may depend on not only their past but also other exogenous variables. E.g. number of bikeshare users may depend on weather.

Modelling time series

Time series display some characteristics unlike usual i.i.d. data.

- Be mindful of autocorrelation in the data
- Consider seasonal and exogenous variables

Cross-sectional and panel data

In cross-sectional and panel data, our observations are correlated as we are taking observations from the same person (across time) or unit.

Example

- Repeated measures across time (follow-up study)
- Study for student performance: students are grouped in classrooms that are grouped in schools.

ANOVA (analysis of variance)

Example

1. Experiment to determine textbook
2. Each class in school is given different textbook
3. Question: is the average performance in each class same or different?

ANOVA (analysis of variance)

Example

1. Experiment to determine textbook
2. Each class in school is given different textbook
3. Question: is the average performance in each class same or different?

This is analysis of variance – equivalent to linear regression.

Mixed effect models

Suppose the textbooks were picked at random, and wish to understand how well student may do with next textbook.

Mixed effects

Model variability in the textbook effect: are they all the same or all very different?

Mixed effect models

Can be used to model shared variability among observations.

- Variability among subjects for experiments with repeated observations
- Variability among units for experiments with several units (e.g. classrooms, schools, etc.)
- Can combine those: i.e. follow students throughout their school years – might change classes.