MONASH University

ETC3250 Business Analytics: Data Wrangling

Souhaib Ben Taieb, Di Cook, Rob Hyndman

September 10, 2015

Handling missing values

- Need to know how the missings are coded, hopefully clearly missing, treated as NA in R, not 0, or -9, or -9999, or . Recode as need be.
- Study the distribution of missing vs not missing, which will help determine how to handle them.

What ways can these affect analysis?

- If missings happen when conditions are special, eg sensor tends to stop when temperature drops below 3 degrees Celsius, estimation of model parameters may not reflect the population parameters
- Some techniques, particularly multivariate methods like many used in data mining require complete records over many variables. Just a few missing numbers can mean a lot of cases that cannot be used.

Terminology

- missing completely at random (MCAR) means that values that are missing appaear to be independent of everything else, just sporadically occur
- missing at random (MAR) means that missings can dependent on other known information, eg temperature, and this information can be used to help estimate values to substitute the missing values
- missing not at random (MNAR) means that the missings are dependent on something else, but we may not have that information, which makes it impossible to appropriately estimate substitute values.

Making it Easy - MissingDataGUI

- Methods for summarising missings in a data set
- Ways to plot to examine dependence between missing vs not missing
- Imputation methods to substitute missings