Methods for dealing with missing values in datasets

Andrew Beveridge - ab441 / H00013703

October 19, 2013

Listwise deletion, imputation (preferably multiple) and interpolation are some methods of dealing with missing values. Listwise deletion of all cases which have missing values is a quick and dirty way to solve the problem, but has many pitfalls, such as introducing bias. The most basic interpolation; linear interpolation, takes the two nearest data points and fills the missing point with a value calculated based on the distance from the case with the missing value to each neighbour. Imputation means replacing missing data with probable values. One method; mean imputation, replaces any missing value with the mean of all other cases of that variable. Multiple imputation (creating several imputed data sets then averaging the outcomes) provides much greater accuracy.

References

- http://j.mp/AderHJ
  (http://books.google.co.uk/books?id=LCnOj4ZFyjkC&lpg=PP1&pg=PA305#v=onepage&q&f=false)
- http://j.mp/MultipleImputation
  (http://books.google.co.uk/books?id=cNvTIOLs_WMC&lpg=PP1&pg=PA15#v=onepage&q&f=false)