Methods for dealing with missing values in datasets

Author, Amer Al Mazloum

A Data Science Foundation Blog

September 2017

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3670 Email:admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

Methods for dealing with missing values in datasets

AlMazloum, Amer Eddin
HERIOT WATT
Professor: Dr. Hani Ragab

- ## Missing Values in Data:

  Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit ("subject"). Some items are more likely to generate a nonresponse than others.

- ## Missing data mechanisms:

  - ### Missing completely at random (MCAR):

    Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. On another hand, missing value (y) neither depends on x nor y.

  - ### Missing at random (MAR):

    The probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (say X). on another hand, Missing value (y) depends on x, but not y.

  - ### Not missing at random (NMAR):

    Missing values do depend on unobserved values. On another hand, the probability

of a missing value depends on the variable that is missing.

- **Patterns of Missingness:**

  We can distinguish between two main patterns of missingness. On the one hand, data are missing monotone if we can observe a pattern among the missing values. Note that it may be necessary to reorder variables and/or individuals. On the other hand, data are missing arbitrarily if there is not a way to order the variables to observe a clear pattern (SAS Institute, 2005).

- **Methods for handling missing data:**

  - **Deletion Methods**

    - **List wise deletion**

      If a case has missing data for any of the variables, then simply exclude that case from the analysis. It is usually the default in statistical packages. (Briggs et al.,2003). In this case, rows containing missing variables are deleted

    - **Pair wise deletion**

      Analysis with all cases in which the variables of interest are present. On another hand, only the missing observations are ignored and analysis is done on variables present.

  - **Imputation Methods**

    - **Popular Averaging Techniques:**

      Mean, median and mode are the most popular averaging techniques, which are used to infer missing values. Approaches ranging from global average for the variable to averages based on groups are usually considered. On simply way Replace missing value with sample mean or mode.

- **Conditional mean imputation:**

  Suppose we are estimating a regression model with multiple independent variables. One of them, X, has missing values. We select those cases with complete information and regress X on all the other independent variables. Then, we use the estimated equation to predict X for those cases it is missing. (Graham, 2009) (Allison, 2001) and (Briggs et al., 2003).

- **Model-Based Methods:**

  - **Maximum Likelihood:**

    We can use this method to get the variance-covariance matrix for the variables in the model based on all the available data points, and then use the obtained variance- covariance matrix to estimate our regression model (Schafer, 1997). On another hand, Estimate: value that is most likely to have resulted in the observed data.

  - **Multiple imputation:**

    The imputed values are draws from a distribution, so they inherently contain some variation. Thus, multiple imputation (MI) solves the limitations of single imputation by introducing an additional form of error based on variation in the parameter estimates across the imputation, which is called "between imputation error". It replaces each missing item with two or more acceptable values, representing a distribution of possibilities (Allison, 2001).

- **How do you deal with missing values?**

  Ignore or treat them? The answer would depend on the percentage of those missing values in the dataset, the variables affected by missing values, whether those missing values are a part of dependent or the independent variables, etc. Missing Value treatment becomes important since the data insights or the performance of your predictive model could be impacted if the missing values are not appropriately handled. **In conclusion:** Assumptions and patterns of missingness are used to determine which methods can be used to deal with missing data

- **Sources and useful resources:**

  - **Reports:**

    - http://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf
    - https://liberalarts.utexas.edu/prc/_files/cs/Missing-Data.pdf

  - **References:**

    - Allison, P., 2001. Missing data — Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136.
    - Enders, Craig. 2010.Applied Missing Data Analysis.
    - STATA 11, 2009, Multiple Imputation. Stata Corp.
    - Schafer, J. L. ,1997. Analysis of Incomplete Multivariate Data.

  - **Useful links:**

  - Data sets with missing values that can be downloaded in different formats including SAS, STATA, SPSS and S plus: http://www.ats.ucla.edu/stat/examples/md/default.htm.
  - Introduction to missing data with useful examples in SAS http://www.ats.ucla.edu/stat/sas/modules/missing.htm.
  - Multiple imputation in SAS. Comprehensive explanations http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part1.htm

Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3670 Email:admin@datascience.foundation Web: www.datascience.foundation

Registered in England and Wales 4th June 2015, Registered Number 9624670

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email: admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation