Missing Data - Introductory notes

David Pham

Contents

1 Wikipedia 1

2 Matloff Blog post 1

3 Gelman, Hill, Data analysis using regression and multilevel/hierarchical models 3

4 Little, Robin, Statistical analysis with missing data 3

Bibliography 3

These are my current notes and the themes

1 Wikipedia

Case deletion (CC) The method does not introduce any bias if the missing values are uniformly distributed.

Single imputation If one sort the data matrix according some order, *Last observation carried forward* is the method of replacing the missing value with last valid value.

The missing value can also be replaced with the mean of the other observations, however, correlations are attenuated.

Regression imputation use the other variables as predictors to replace the missing value, although precision is misleadingly augmented, hence does not reflect the statistical errors of the missing data. This problem is partially solved by multiple imputation.

Multiple imputation The multiple imputation [Rubin (1987)] is similar to bootstrapping method: Missing variables are simulated, say B times, and the desired statistics are averaged except for the standard error which is constructed by adding the variance of the imputed data and the within variance of each data set.

2 Matloff Blog post

Complete-case analysis (CC), listwise deletion Delete all record for which at least one variable is missing.

Single and multiple imputation Estimation of the distribution of missing variables conditional on the others and then sampling from that distirbution. Multiple alternate matrix are generated without the NAs.

In multiple imputation, the distribution of each variable conditional and the others is fitted and in case of missing value, a sample is drawn from this distribution.

Available cases (AC), pairwise deletion Keep the observation if the missing feature is not retained for the desired measure, for example the correlation (where only 2 variables are needed). It can, nonetheless, produce correlations over 1.

2.1 MCAR: Missing Completely at Random

Let Y the variable of interest, $M \in \{0,1\}$ denotes if Y is missing, and D the other variables than Y. This is often denoted as

$$P(M = 1|Y = s, D = t) = P(M = 1),$$

or equivalently

$$P(Y = s, D = t|M = i) = P(Y = s, D = t), i \in \{0, 1\}.$$

2.2 MAR: Missing at Random

For multiple imputation, one requires only $M \perp Y|D$, that is

$$P(M = 1|Y = s, D = t) = P(M = 1|D = t),$$

2.2.1 Conditional estimation under MAR

In practice, problems arise as D might not hold any predictive ability of the desired variable and that D might as well contain missing data. Interestingly

$$\begin{aligned} P(Y = s|D = t, M = i) &= \frac{P(Y = s, D = t, M = i)}{P(D = t, M = i)} \\ &= \frac{P(Y = s, D = t)P(M = i|Y = s, D = t)}{P(D = t, M = i)} \\ &= \frac{P(Y = s|D = t)P(D = t)P(M = i|D = t)}{P(D = t, M = i)} \\ &= P(Y = s|t). \end{aligned}$$

Hence if we are interested in the relationship between Y and D, that is the conditional distribution Y given D, the fact that it is missing or not will not introduce bias, hence CC and AC would perform equally well. This is ironic as MAR is meant to apply where CC and AC should not be used.

2.2.2 Unconditional estimation under MAR

Observe that

$$P(Y = s|M = 0) = \frac{P(M = 0|Y = s)}{P(M = 0)} P(Y = s),$$

hence our estimation of $P(Y = s)$ might still be biased with the factor of $P(M = 0|Y = s)/P(M = 0)$.

3 Gelman, Hill, Data analysis using regression and multilevel/hierarchical models

Chapter 25 of (Gelman and Hill 2006) contains information about missing values.

4 Little, Robin, Statistical analysis with missing data

The monograph (Little and Rubin 2002) describes mechanisms underlying the missingness come in several type (*mi*, *mice*, *Amelia* in R packages).

Bibliography

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/hierarchical Models.* Cambridge University Press.

Little, RJA, and DB Rubin. 2002. "Statistical Analysis with Missing Data." Wiley.