GUIDELINES FOR MULTIPLE IMPUTATIONS IN REPEATED MEASUREMENTS WITH TIME-DEPENDENT COVARIATES: A CASE STUDY

Phitchaya Faramnuayphol

IMPUTATION

- Methods that handle missing values depend upon the mechanism of missingness. Missing data mechanism refers to the underlying process of generating missing data.
- It is still considered as imputed data, not real data.
- The rule of thumb suggests that 20% or less of missing data is acceptable rate to use imputation methods

MISSING DATA

- based on the occurrence in time

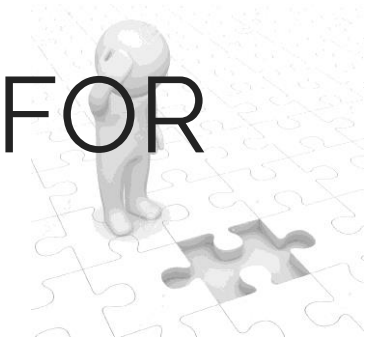
Intermittent pattern (nonmonotone)

missing values due to occasionally omission, with observed values afterwards.

Dropout pattern (monotone)

missing values due to premature withdrawal, with no observed values afterwards

3 MAIN APPROACHES TO COMPENSATE FOR MISSING VALUES



- 1. Imputation using zero, mean, median or most frequent values
- 2. Imputation using random values
- 3. Imputation with a models
  - Model-based imputation with uncertainty
  - Model-based progressive imputation

International Journal of Statistical Distributions and Applications
2017; 3(4): 72-80
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20170304.13
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

SciencePG
Science Publishing Group

Imputation Methods for Longitudinal Data: A Comparative Study

Ahmed Mahmoud Gad[1], Rania Hassan Mohamed Abdelkhalek[2]

[1]Statistics Department, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt

[2]Department Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University, Benha, Egypt

Email address:
Ahmed.gad@feps.edu.eg (A. M. Gad), gendy176@yahoo.com (R. H. M. Abdelkhalek)

To cite this article:
Ahmed Mahmoud Gad, Rania Hassan Mohamed Abdelkhalek. Imputation Methods for Longitudinal Data: A Comparative Study. *International Journal of Statistical Distributions and Applications*. Vol. 3, No. 4, 2017, pp. 72-80. doi: 10.11648/j.ijsd.20170304.13

Received: March 5, 2017; **Accepted**: March 28, 2017; **Published**: November 10, 2017

THE MEAN SUBSTITUTION METHOD (MS)

- The mean value of non-missing observations is used to fill in missing values for all observations.
- maintains the same sample size from reduction
- Can distort the distribution of the variable in large missingness rate
- The covariance also is underestimated because the mean imputation for the missing subjects has zero variance.
- similar to the CCA; it requires MCAR assumption

THE LAST OBSERVATION CARRIED FORWARD METHO (LOCF) (1)

- This method imputes the unobserved value by the last observed value for the same subject.
- For dropout missingness, last observation remains the same after dropout.
- In longitudinal data, it tends to underestimate the true variability of the data.

THE LAST OBSERVATION CARRIED FORWARD METHO (LOCF) (2)

- Advantage
  - can give satisfactory results, if the observations in the dataset are approximately close to each other.
  - When the measurements occasions are short to some extent, this ensures the effectiveness of the LOCF method
- Disadvantage
  - creates bias even if the strong MCAR assumption is satisfied.

THE HOT DECK (HOT) METHOD (1)

- Replaced by a similar responding unit in the same sample. (the most similar subject is selected) or based on the correlation
- Performs well; the variable is highly predictive of missing values and large sample
- It preserves some of the measurement error.

THE HOT DECK (HOT) METHOD (2)

- Cautions; distorting of both correlations and covariance because the missing values are replaced with values that already exist in the distribution of scores.
- The smaller standard errors lead to greater likelihood of a Type I error

K-NEAREST NEIGHBOR (K-NN)

- Appropriate only when it’s MCAR.
- Each imputed value is selected from the respondent who is the nearest to the subject with missing value based on the distance between them.
- The distance is computed using the information from the observed data.
- provides asymptotically valid distribution.
- normally point estimates with small or negligible bias, assuming that a linear relationship exists between the variable of interest y and the concomitant variable x used for nearest neighbor identified.

THE EXPECTATION MAXIMIZATION (EM) ALGORITHM (1)

- finds the parameters which maximize the log-likelihood function when there are missing values in the dataset.
- Two steps:
  - the expectation step (E-step) : calculates the conditional expectation of the complete data log-likelihood

$$Q\left(\theta \mid \hat{\theta}\right) = E\left[g\left(\theta \mid Y\right) \mid Y_{obs}, \theta = \hat{\theta} \right]$$

$$= \int g\left(\theta \mid Y\right) f\left(Y_{mis} \mid Y_{obs}, \theta = \hat{\theta}\right) dY_{mis}$$

  - where is an estimate for θ and g (θ|Y) is the complete data log-likelihood

THE EXPECTATION MAXIMIZATION (EM) ALGORITHM (2)

- the maximization step (M-step); maximize the complete data loglikelihood produced from the E-step to obtain updated parameter estimates.
- The M- step can be expressed as follows:

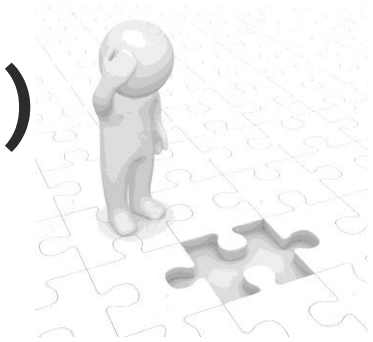
$$Q\left(\theta^{t+1} \mid \hat{\theta}\right) \geq Q\left(\theta \mid \hat{\theta}\right) \text{ for all } \theta$$

- The iteration are continued until some convergence is met

THE EXPECTATION MAXIMIZATION (EM) ALGORITHM (3)

- **Advantages.**
  - the observed data likelihood increases at every step.
  - it is preferred to regression imputation because the estimated parameter values that maximize the observed data log-likelihood function are consistent, efficient under MAR condition and tend to be approximately unbiased in large samples and normally distributed.
  - the obtained variances are close to what is theoretically desirable.
- **Disadvantage** ; the convergence of the iterations can be very slow

THE REGRESSION IMPUTATION (REGRESS) METHOD (1)



- Identifying several predictors for the variable with missing values using a correlation matrix.
- The best predictors used as independent variables. Missing data is used as a dependent variable.
- This variable is regressed on all other variables with complete data for the predictor variables.
- The regression equation is used to replace missing values for incomplete subjects with the predicted values.

THE REGRESSION IMPUTATION (REGRESS) METHOD (2)

- In an iterative process, the values for the missing variable are inserted and then all subjects are used to predict the dependent variable.
- the last round; the predictors are used to replace the missing values.
- Subjects with the same covariates will exactly have the same imputed value

THE REGRESSION IMPUTATION (REGRESS) METHOD (3)

- Advantage
  - regression parameter estimates based on regression imputation under MCAR are relatively unbiased in large samples.
- Disadvantage
  - tend to fit a regression line together too well.
  - they do not reflect the random error or variance (lead to small standard errors and p-values at the time of analysis.)
  - the correlations with the imputed variables are overestimated

MULTIPLE IMPUTATION (1)

- involves imputing each missing value by two or more acceptable values to produce several different complete datasets.
- Then each dataset is analyzed to produce different parameter estimates. (each imputation are then combined using a special rule)
- A key feature is that the uncertainty about the parameters in the imputation model is taken into account when imputing the unobserved values.

MULTIPLE IMPUTATION (2)

- Advantage
  - less biased compared with the estimates obtained from single imputation methods.
  - provides more correct standard errors, P-values, and confidence intervals as opposed to single imputation methods, which gives too small standard errors.
- Disadvantages
  - The uncertainty inherent in missing values is ignored
  - Procedure takes more work both to create the imputations and to analyze the results
  - Does not satisfy normality test in most situations.

RESULTS

Methods	Advantage	Disadvantage
CCA	The first choice of imputation even in MCAR in MCAR, it has the least bias.	In MAR and MNAR, it gives biased estimates but small SME.
Regression method	performs well especially under the MCAR	but the sample variance and covariance are underestimated which leads to small standard error and P-value.
Mean imputation method	It performs slightly well for the MAR and the MNAR assumptions and produced less MSE compared to other methods.	is not a good choice for the dropout pattern under the MCAR assumption.
The LOCF	estimates the parameter very well and gives small MSE except under MNAR assumption.	large bias in some parameters.
The HOT method	sustains from large bias especially with MNAR missingness. It's performance gets better for large samples under the MCAR and the MNAR.	However, the HOT has small MSE under the three missingness mechanisms.

RESULTS

Methods	Advantage	Disadvantage
KNN	The gives reasonable results for the MCAR and the MAR mechanisms.	It gets better results as the sample size increase, in other word it should be applied for large sample sizes rather than small sample sizes.
EM algorithm	However, it gives small MSE compared to the other methods.	The provides a poor prediction to missing values under the three missing data mechanisms especially the MCAR.
MI method	The estimates are relatively biased, but under the MCAR mechanism it has the least bias.	provides small MSE.

MULTILEVEL IMPUTATION

- Multilevel data
  - a hierarchical or nested data structure
  - consist of repeated measures within subjects, or respondents within clusters, as in cluster sampling.
- Imputation methods
  - Joint modeling approach (JM)
  - Fully conditional specification of MI (FCS)

JOINT MODELING APPROACH (JM) (1)

- a single model is specified for all variables with missing data, and imputations are simultaneously generated from this model for all variables with missing data.
- For individual-level variables

Y_{ij} contains a number of individual-level target variables with arbitrary patterns of missing data

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_j + \mathbf{e}_{ij},$$

e_{ij} denotes the residuals

X_{ij} contains fully observed predictor variables with associated fixed effects β

Z_{ij} contains fully observed predictor variables with associated random effects uj

JOINT MODELING APPROACH (JM) (2)

- Assume that X and/or Y are partially missing. Treating both X and Y as target variables

$$[X_{ij}, Y_{ij}]^T = [\beta_{0(x)}, \beta_{0(y)}]^T + [u_{j(x)}, u_{j(y)}]^T + [e_{ij(x)}, e_{ij(y)}]^T,$$

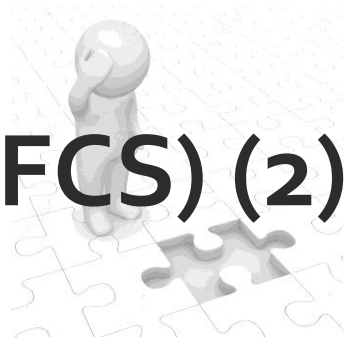
FULLY CONDITIONAL SPECIFICATION OF MI (FCS)(1)

- imputes missing data separately for each variable with missing data, conditioning on some or all of the other variables in the data set.
- iterates back and forth between different target variables.

$$X_{ij} = \beta_{0(x)} + \beta_{1(x)}(Y_{ij} - \bar{Y}_{\bullet j}) + \beta_{2(x)}\bar{Y}_{\bullet j} + u_{j(x)} + e_{ij(x)}$$
$$Y_{ij} = \beta_{0(y)} + \beta_{1(y)}(X_{ij} - \bar{X}_{\bullet j}) + \beta_{2(y)}\bar{X}_{\bullet j} + u_{j(y)} + e_{ij(y)}.$$

- If both variables are affected by missing data, then the group means are updated at each iteration of the sampling algorithm on the basis of the most recent imputations for X and Y.

FULLY CONDITIONAL SPECIFICATION OF MI (FCS) (2)



- FCS approach relies on the observed group means to represent the different relations between X and Y at the individual and the group level.
- For each additional variable with missing data, an additional equation must be specified, each conditioning on the other variables and their respective group means

THANK YOU