African Journal of Applied Statistics, home page : www.jafristatap.net
Vol. 2, 2015, pages 29–37.
DOI : http ://dx.doi.org/10.16929/ajas/2015.1.29.73

ISSN 2316-0861

ANALYSIS OF MISSING DATA IN SERO-EPIDEMIOLOGICAL STUDIES

[1,2]**Oumy Niass,** [1]**Abdou Kâ Diongue** [1], [2]**Aissatou Toure**

[1] *LERSTAD : Laboratoire d'Etudes et de Recherches en Statistiques et Développement, UFR Sciences Appliquées et Technologie, UGB, BP 234, Saint-Louis–Sénégal*
[2] *Unité d'Immunologie, Institut Pasteur de Dakar, 36 Avenue Pasteur, BP 220 Dakar–Sénégal*

Presented by Professor Gane Samb LO, member of the Corresponding Editors Board

Résumé. **(French)** L'analyse des données manquantes est un problème récurrent dans les études biologiques en particulier dans les études séro-épidémiologiques. La méthode la plus couramment utilisée est celle dite de restriction qui consiste à restreindre l'analyse sur les individus ayant des informations complètes sur toutes les variables de la base de données. Cette méthode pourrait entrainer des pertes d'information ou introduire des erreurs sur l'évaluation des résultats. Le but de cette étude est de comparer des techniques d'évaluation des données manquantes et de démontrer que l'estimation des données manquantes est parfois plus efficient que la suppression. Nous utilisons des données transversales collectées sur 300 enfants vivant dans huit villages dans le but de mieux comprendre la relation entre les réponses d'anticorps dirigées contre les différents antigènes du Paludisme. La base complète a été utilisée pour créer des bases incomplètes avec des pourcentages de données manquantes variant de 5 % à 50%. Les six méthodes suivantes : méthode des cas complète (CC : méthode par suppression), méthode de substitution par la moyenne, méthode des plus proches voisins (knn), méthode de l'imputation multiple par l'algorithme EM, la méthode du predictive mean matching (pmm) et méthode de la régression, ont été appliquées sur dix bases incomplètes sur le jeu de données décrit plus haut. Les indicateurs statistiques suivants ont été utilisés pour comparer ces différentes méthodes : erreur quadratique moyenne, erreur absolue moyenne, niveau de signification (p-values), sommes des erreurs quadratiques, critères AIC et BIC. Les résultats montrent que lorsque le pourcentage de données manquantes est supérieur à 5%, la méthode MI.pmm et celle des plus proches voisins donnent les meilleurs résultats. Lorsque le pourcentage de données manquantes est supérieur à 5%, la méthode MI.pmm et celle des plus proches voisins donnent les meilleurs résultats. La méthode par suppression s'est révélée comme étant la plus inappropriée. En se basant sur les résultats, il s'avère qu'il est préférable d'estimer les données manquantes que de les supprimer.

Abstract (English) The treatment of missing data represents a recurrent problem in biology, in particular in the sero- epidemiological studies. Indeed, the most common method used to deal with missing data is to restrict the analysis to subjects having complete information for the set of variables of interest, which can lead to a drop-out and/or introduce some slants in the evaluation. The aim of this paper is to compare some missing data techniques and demonstrate that estimating missing data is sometimes more efficient than deleting them. Cross-sectional data was obtained by investigating the relationship between different malaria antibody responses against some antigens of *P.falciparum* in a sample of 300 children from eight villages in a rural area of Senegal (West Africa). The complete dataset was used to create incomplete dataset with percentages of missing values varying between 5 % to 50%. Six methods were tested for dealing with missing values : Complete-case (CC) analysis so-called listwise deletion, mean substitution, k-nearest neighbours (knn), multiple imputation using the expectation-maximization (EM), predictive mean matching (pmm) and regression. They were applied to ten incomplete dataset for the same missing position. Root mean square errors (RMSE), mean absolute errors (MAE), p.value, multiple R-square, AIC and BIC criterions were used to compare these missing data approaches. The results demonstrate that multiple imputation using predictive mean matching (MI.pmm) and k-nearest neighbors (knn) methods were preferable to other missing data ones when the missing data percentage was great (larger than 5 percent). The listwise deletion approach produces the most inaccurate results. Based on these results, it seems that it is preferable to estimate missing values than to restrict the analysis to the subjects who have complete observations.

Keywords : missing data, imputation, Plasmodium falciparum, serology.
AMS 2010 Subject Classification : 62P10.

Article history : Received 2015/10/08 ; Accepted 2015/12/15 ; Published Online 2015/12/31.

1. Introduction

The results of serological studies are extensively based on the findings from statistical analysis of collected data. However, all observations are not always informed. The handling of missing data is a sensitive issue as the data processing management can affect the results of the analysis or the parameters of interest. The missing values can be in the variable of interest and/or in the independent variables. The reason for which a

1. Corresponding author : (Abdou Kâ Diongue) : abdou.diongue@ugb.edu.sn
Oumy Niass (niass_oumy@yahoo.fr, oumyniass@gmail.com),
Abdou Kâ Diongue (abdou.diongue@ugb.edu.sn),
Aissatou Toure (atoure@pasteur.sn)

measure was missing are numerous. For example, a subject may refuse to participate completely in the study, or can miss at the moment of blood sampling (Cross-sectional study). In a longitudinal study, subject may drop out or be absent by administrative or personal activities during the sampling. There are several techniques to manage the problem of incomplete data, going from restricting the analysis to units that have complete measures for all the variables in the set to the replacement of the incomplete data by plausible(s) value(s). However, some methods lead to inefficient analysis and commonly produce highly biased estimates in the association of the variables studied (see Greenland S. and Finkle (1995), Little (1992)).

The goal of this paper is to compare some estimated results of different imputation methods dealing with the problem of missing values. We used a real database in which we applied the so-called simple imputation method that consists in replacing missing values by a single value, and the methods of multiple imputation. The paper proceeds as follows. In Section 2, we discuss the mechanism of missing data. In Section 3, we introduce some methods for dealing with missing data. In Section 4, we explain the methodology adopted in this study. In Section 5, we describe the utility of each missing data technique in the context of cross-sectional data analysis, and in the final section we conclude with a discussion.

2. Mechanism of missing data

When managing missing data, it is helpful to know the mechanism of missingness. That is, the reason for why data are missing. The data set is represented in general by a table whose lines represent the subjects and whose columns consists of the measures of the variables pertaining to the subjects. Missing data are the unobserved data. They are represented by the symbols NA (Not Available) as in Table 6.1.

In the literature, there are three distinct types of non-response mechanism (see Rubin (1976)).

2.1. **Missing Complement At Random (MCAR).** A variable is missing at random if the probability that an observation is missing is independent for any characteristics of the units. That is to say, the probability of missing for a given variable does not depend on it, but only on external parameters independent from this variable. For example, antibodies measurements may be missing because ; the tube containing blood sample of subject is broken by accident or by malfunctioning materials of laboratories during handling. When missingness are MCAR, most of simple techniques for managing missing data, give unbiased results (see Greenland S. and Finkle (1995)).

2.2. **Missing At Random (MAR).** Data is missing at random, if the probability of non-response depends only on the observed data. For example, younger people might be more likely to miss antibody responses measurement than older people. An antibody response measurement is MAR, if the study has collected information on age for all the subjects in the survey.

2.3. **Missing Not At Random (MNAR).** Data are MNAR, when the probability that a data is missing depends both on the observed and on the missing data.

The statistic management of missing data is greatly based on the understanding of missing value mechanism. The MCAR and MAR contexts are easiest to solve because the observed data contain all the necessary information to estimate the missing data distribution. But the MNAR is the situation which is most problematic because it leads to biased estimates (Yulei (2010)).

3. Methods for handling missing data

3.1. **Complete-Case analysis (CC analysis).** The standard method is to restrict the analysis to units with no missing values for all the variables in the set. This option referred to as Complete-Case analysis or the listwise deletion analysis consists in eliminating any case with missing measurements for any variables used in the undergoing analysis. It is the default technique in many statistics softwares and can biase the results (Raghunathan (2014)). If the data are MNAR, the complete case analysis can introduce a systematic bias defined by the behavior of missingness. But if data are MAR or MCAR, the listwise deletion analysis will lead to a reduced sample size and to lesser powerfullness to detect statistical effects (Alison (2002)). Nevertheless, the listwise deletion analysis is reasonable if the fraction of missing data in the set is at most 5% (Yulei (2010)).

3.2. **Simple imputation approaches.** In this section, we present some simple imputation approaches like the mean substitution, the regression substitution, the k-nearest neighbours methods as well as the expectation-maximization algorithm allowing to handle the problem of missing data.

3.2.1. *Mean substitution.* The mean substitution is a technique allowing to treat missing data that consists in substitution for a given variable, each missing value by the mean of the observed values. This approach preserves the mean of the variable distribution but reduces other characteristics of the variable distribution (Rubin (1987) , Cole (2008)) . Allison showed that mean substitution approach restricts the variability of a variable and changes the underlying distribution (Alison (2002)). These distributional problems are the reason why statisticians are more lead to suggest the complete-case analysis instead of the mean substitution analysis (Little (1989)) .

3.2.2. *Regression substitution.* The principle of regression method is to use the observed values to create fitted regression model. The variables with missing data are the variables of interest and missing values are replaced by the predicted values according to the model.

Explicitly, we suppose that X is a matrix which represents the dataset, $X = (X^{(1)}, X^{(2)}, \cdots, X^{(p)})$ where each column $X^{(i)} (i = 1, 2, \cdots, p)$ is a random variable of n lines. Let $X^{(j)}$ a column with missing values. Set $X^{(j)} = (X^{(j)}_{obs}, X^{(j)}_{miss})$ where $X^{(j)}_{obs}$ is the sub-vector of observed values of $X^{(j)}$ and $X^{(j)}_{miss}$ that of the missing values. We consider $H = \{i, X^{(i)}_{miss} = \phi, i = 1, 2, \cdots, p\}$ with cardinality m, the set of indices of columns which not having missing values and $Z = \{X^{(i)}, i \in H\}$. Let Z_{obs} and Z_{miss}, two sub-matrix of Z extracted by selection of lines corresponding respectively to $X^{(j)}_{obs}$ and $X^{(j)}_{miss}$. Let's consider the regression model based in the observed part :

$$X_{obs} = \beta Z_{obs} + \mu, \qquad \mu \rightsquigarrow N(0, \sigma), \tag{3.1}$$

with $\beta = (\beta_0, \beta_2, \cdots, \beta_m)$ is the vector of regression coefficient parameters and the error $\mu = (\mu_1, \mu_2, \cdots, \mu_n - q)$ where q is the length of $X^{(j)}_{miss}$.

The estimation of the missing values, $\hat{X}^{(j)}_{miss,i}$ where i ranges over the q line indices of $X^{(j)}_{miss}$, are obtained by

$$\hat{X}^{(j)}_{miss,i} = \hat{\beta}_0 + \hat{\beta}_1 Z_{miss,1} + ... + \hat{\beta}_m Z_{miss,m},$$

where $\hat{\beta}$ is the usual estimator of β.

The regression approach when dealing with missing values depends on the predictors that are considered into the equation regression method. It is the reason why Little (2002) consider that this technique is a conditional one. It is more sophisticated than mean substitution method Rubin *et al.* (2007), but this technique can conduct to overestimating the relationship between the predictors and the dependent variables (Schafer and Graham (2002)).

3.2.3. *The nearest neighbours Imputation (Knn).* The nearest neighbours imputation method is a technique based on the notion of proximity between observations (subjects). This similarity was often determined by a distance function (the Euclidian distance, for example). It is a technique according to which the missing data for a given subject are replaced by the value observed at the same position of the nearest subject. According to the previous notations in Section 3.2.2. Let Z_{obs} and Z_{miss}, two sub-matrix of Z extracted by the selection of lines corresponding respectively to $X^{(j)}_{obs}$ and $X^{(j)}_{miss}$. We assume that l is the identifyer of the subject who has not observed value for the variable $X^{(j)}$. We let among the subject k who has all measurement in the set, and the subject j_0 who minimizes the distance between k and l :

$$j_0 = \arg \min_{1 \leq k \leq n} d(Z^{(i)}_{obs}(l), Z^{(i)}_{obs}(k)), i \in H \tag{3.2}$$

with d a distance measure and n = number of subject in the set. Here d is the Euclidian distance defined by

$$d\left(Z^{(i)}_{obs}(l), Z^{(i)}_{obs}(k)\right) = \sqrt{\sum_{i \in H} \left(Z^{(i)}_{obs}(l) - Z^{(i)}_{obs}(k)\right)^2}$$

If j_0 was determined missing value, $X^{(j)}_{miss}(l)$ would be estimated by $X_{obs}(j_0)$: $\hat{X}^{(j)}_{miss}(l) = X^{(j)}_{obs}(j_0)$.

3.2.4. *Expectation-Maximisation.* The EM algorithm, initially developed by Dempster *et al.* (1977), is an iterative algorithm for maximizing the likelihood estimated by a parametric model for observed data. EM missing method is based on the maximum-likelihood estimated of the covariance structure given by the available data. The EM is a succession of two state steps. In the Expectation state step (the first step), regression equations based on the given observed data are used to estimate missingness ("the expected values"). These missingness are replaced by the conditional mean based on the regression equations. In the maximization state step the estimates obtained from the expectation state are updated to maximize the log likelihood of the current parameters from the first state. These two steps are repeated for some number of iterations. This algorithm will converge on a stationary point under some hypotheses of regularity (Alison (2002), Dempster *et al.* (1977)).

3.3. **Multiple Imputation (MI).** The MI technique approach was proposed Rubin (1978), and by Rubin (1987), and by Schafer (1997). The MI method replaced missing values by more than one value given by statistical models, generating more completed data set (Figure 6.1). The MI approaches are methods that use a variety of advanced techniques of imputation to estimate missingness, creating more versions of the same data set. Their aim is to correct the under-estimate of the variance that is a characteristic of the single imputation. The MI methods add a correcting factor of the variance calculated from the interimputation in order to take into account the uncertainty of the missing data estimates. Rubin (1996) described the MI method as a succession of three states step. First m values ($m > 1$) are assigned to the missing data, generating thus m completed data sets. Secondly, these completed data sets are separately analyzed using standard statistical methods like regression approaches, multivariate models, ANOVA anaysis, etc· Finally, the results are combined to produce estimates and confidence intervals. These results are more robust than those given by simple imputation (Schafer (1997), Schafer (1999)).
In this paper, we consider this technique by using the predictive mean matching method, the EM algorithm, and the regression methods described in previous sections.
The predictive mean matching scheme is similar to the regression method except the fact that for each missing data, a value is randomly imported from a set of observed values whose predicted values are the closest to the predicted value for the missing value from the simulated regression model ((Zio Di and Guarnera (2009)).

4. Materials and Methods

4.1. **Data collection.** A total of 300 observations collected from children living from eight villages located in a rural area of Senegal were included in this study. These villages were selected for their proximity to Dielmo and Ndiop villages which main epidemiological features have been reported previously in Rogier *et al.* (1999) and in Roussilhon *et al.* (2007). After the informed consent from each legal represent was obtained, venous blood sample were collected during the lowest transmission season in both villages. These blood samples were transferred to a laboratory in Dakar and used for serological studies. In this paper we use antibody responses against four antigens of *P. falciparum*. The antibody level estimations have been obtained by ELISA (Enzyme Linked immunosorbent Assay), previously described in Toure et *al.* (2009).

4.2. **Data analysis.** With the aim to compare different approaches for handling missing data, we have extracted, the immunological data from the database in a fisrt time. These data represents our matrix of reference (MR) in which there are not missing values. To simulate the MCAR situation, τ% of the matrix of reference (MR) were randomly selected to be identified as missing. The number τ is the percentage of missing data and it is varying between 5 and 50 percent. Thus we create ten matrix with missing values (MVM) containing respectively 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 percent of missing data in the same matrix of reference (MR). For every MVM, each approach is applied to the exact same missing value point. **The statistics Analysis software R**, version 2.15.1, is used for all analysis. The "amelia" package is used for the multiple imputation with the Expectation-Maximisation (MI.EM), "kNN" package for the nearest neighbours imputation, and "mice" package for multiple imputation with predictive mean matching (pmm). Three criterions are used to compare imputation methods : Mean Absolute Error (MAE), Root Mean Square Error (RMSE) given by the equations (4.3 and 4.4) and mean residuals.

$$RMSE_\tau = \sqrt{\frac{\sum_{i=1}^{M_\tau} (R_i - E_i)^2}{M_\tau}} \tag{4.3}$$

$$MAE_\tau = \frac{\sum_{i=1}^{M_\tau} \|R_i - E_i\|}{M_\tau} \tag{4.4}$$

where τ is the percent of missing data, M_τ the number of missingness in the set, R_i is the real value at the position i and E_i the estimate value at the same position by missing imputation methods.

4.2.1. *Multiple linear regression with missing data.* The multiple linear regression was used to find the relationship between antibody responses against two P. falciparum related variables and two recombinant antigens (AMA1 and GLURP). The two P. falciparum ralated variables are two crude extract strains : the variable Palo Alto which is the reference, and the local strain adapted to culture in the laboratory. We considered the multiple R-square, AIC and BIC criterions and the p.value to assess the impact of missing data technique in model estimates.

5. Results and Discussion

For a dataset of 300 observations, we have simulated 10 databases with percentages of missing data varying between 5 and 50. Also by simulation and by application of missing data techniques we have created 290 completed datasets that are analyzed in this study. In order to compare missing methods, residual means and standard deviations for antibody responses measurement against P. falciparum antigens are calculated by using the technique described by Rubin *et al.* (2007), for each method that deals with missing data and percentage of missingness. Antibody responses measurement against P. falciparum antigens mean and standard deviation from the original dataset are respectively 2.666 and 1.52. Table 6.2 presents the results of the residuals.

The results show a small difference for mean estimates and standard deviation estimates for all missing data techniques and the percentage of missing data compared to the real mean and real variance (Table 6.2). For the mean estimate the multiple imputation methods give the most accurate results but for variance estimate mean substitution gives the most inaccurate estimate up to 45% of missing values. These results have been shown previously in Melanie and Berchtold (2010). Essentially the listwise deletion method produces the most inaccurate estimate of the mean as it was demonstrated with small sample with 10% missing data (see Rubin *et al.* (2007)).

RMSE's and mean absolute errors (MAE) are calculated by the formula given in equations (4.3 and 4.4) and plotted as function s of missing values techniques s and percentages of missing values for the dataset (Figure 6.2).

Based on these results, we observe that RMS's and mean absolute errors are greater for the mean substitution method than for other missing data methods. We also observed small differences in the estimates values. We expect little variability in the estimate of multiple imputation technique using predictive mean matching (MI.pmm) and the nearest neighbors imputation (I.Knn). The Multiple imputation and nearest neighbors imputation methods are preferred to the mean substitution one.

Figure 6.3 provides the histogram of estimated regression coefficients before and after deletion using Complete-Case analysis. We notice deformation of the distribution tail when the proportion of missing values are larger than 5 pourcent. This result seems confirmed by previous results describing eventual biais expected with the listwise deletion method (Raghunathan (2014)).

The AIC, BIC (Figure 6.4) and multiple R-square statistics that were computed from the multiple regression model are plotted as functions of missing data approach and percentages of missing for the main impact of antibody responses against recombinant antigens (AMA1, MSP3 and GLURP) and against local strain (F15) in antibody responses against the Palo Alto (PA) strain (Figure 6.5). Both AIC and BIC indicate that there are large differences between complete-case analysis (CC) and missing data imputation techniques. Results of the listwise deletions approach in terms of AIC and BIC are less than those from the complete dataset. The mean substitution (I.Mean), the nearest neighbours (I.Knn) methods as well as the multiple imputation using predictive mean matching (MI.pmm) give all efficient results. Therefore when missing data range from 5 percent to 35 percent, the multiple imputation approach using the EM algorithm (MI.EM), and the one using the regression scheme (MI.Reg) are also efficient. But above 35 percent of missing data, they underestimate the results. Under 5 percent of missing data, all missing data methods except the MI.EM give efficient estimate for multiple R-square (Figure 6.5). Over 5 percent of missing data, the mean substitution and the (MI.pmm) approaches underestimate the multiple R-square statistic from the complete dataset but the methods MI.Reg and MI.EM over-estimate it. The listwise deletion results in Multiple R-square are less accurate between 5 percent to 30 percent of missing data and greater over 30 percent of missing values than multiple R-square. From the complete dataset on the contrary the nearest neighbors technique (I.Knn), results are so accurate. The method of mean substitution and That of I.Knn as well as the multiple imputation method using predictive mean matching have advantage over the of methods of MI.EM and MI.Reg in terms of AIC and BIC and while the nearest neighbors method (I.Knn) produces most accurate results in term of multiple R-square. It is necessary to note that above 5 percent, The CC analysis under-estimates the regression parameters and the standard errors. We also observe that some covariables which are significantly contributing become not significant over 15 percent for the listwise deletion sheme, over 20 percent for single imputation method and over 25 percent for multiple imputation approach.

The mean substitution and the I.Knn methods and the multiple imputation method using the predictive mean matching have advantage over The methods of MI.EM and of MI.Reg in terms AIC and BIC while and the nearest neighbors approcah (I.Knn) produces most accurate results in multiple R-square.

It is necessary to notice that above 5 percent the CC approach under-estimates the regression parameters and the standard errors. We observe also that some covariables which are significantly contributing become not significant over 15 percent for the listwise deletion method, over 20 percent for the single imputation approach and over 25 percent for the multiple imputation one.

6. Conclusion

In this paper, we focus on missing data problems and techniques to manage it for cross-sectional analysis. We study methods for dealing with missing data on a complete dataset that examine the profile of antibodies response directed against crude extracts of two strains and four peptides of Plasmodium falciparum malaria parasite using frequencies analysis. Our results show that the deletion method is the least efficient for both mean and variance estimates as well as estimates of regression attributes. With 5 percent of missing data , the methods of mean substitution, of the nearest neighbours and of the predictive mean matching are efficient procedures according to regression expectations (AIC, BIC and R-square) compaired with the multiple imputation method using the EM algorithm and to regression methods. When the percentage of missingness increases to 50 %, the AIC and BIC show that The EM and the regression methods have less advantage over the methods of mean substitution, of the nearest neighbours, and of the predictive mean matching.

Nevertheless, the mean substitution method is the most inaccurate among all these techniques when considering the MAE criteria, but it gives a small advantage compare to the complete analysis approach when looking the p-value.

The analysis shows that when more than 5 percent of missing data is considered, the listwise deletion method is not an effective one. This finding could lead to serious consequences since that most sero-epidemiological researchers tend to rely on suppression method leading to the complete case analysis where subjects having only complete information on all explanatory variables are included in the study. This technique assumes that the mechanism of missing data is MCAR. However, if the dataset obtained after deletion is not representative, then the probability of inaccurate results increases. In addition, the data can be missing at random (MAR) or not at random (MNAR). However our approach needs to be improved. Indeed, we use a real dataset instead of simulated data when comparing the different methods to handle missingness while controlling the distribution of the explanatory variables. However, one weakness of using real data is that the results could be specific to particularities in the data set or in the sampling, or to the theoretical expectation. Then, the results are specifically sero-epidemiological studies using linear regression with mixte data. We do not compare the likelihood and the MCMC methods that seems to be the most preferred because of the requirements of commonly used softwares. Given the above-mentioned limitations, it is necessary to generalize these results to other statistical analysis (mixed effect regression,$\cdots$) and also to consider the case where missing data are MAR or MNAR. With the current progress of analytical tools and advancement in missing data techniques, researchers are able to go beyond complete cases analysis or the mean substitution scheme. The deletion method is unaffected for the multivariate analysis in particular when the percentage of missing data is important.

Acknowledgements

This research was supported by grants from EDCTP and Pasteur Institut. The authors thank all the staff of Immunology unit in particular M. Fode Diop (PHD, student) for his great help in data collection.

Références

Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* ; **142** : pp. 1255e64. 1995.

Little RA. Regression with missing x's. *Journal of American Statistical Association* ; **87** : pp.1227e37. 1992.

Rubin DB. Inference and missing data. *Biometrika* ; **63** : pp. 581–592. 1976.

Raghunathan Trivellore E. What do we do with missing data ? Some options for analysis of incomplete data. *Annual Review of Public Health* ; **25** : pp. 99–117. 2014.

Yulei He. Missing data analysis using multiple imputation getting to the heart of the matter. *Circulationcardiovascular Quality and Outcomes* ; **3** :pp.98–105. 2010.

Allison PD. Missing data. *Thousand Oaks, CA/ Sage Publication*. 2002.

Rubin DB. Multiple imputation for non response in surveys. *New York : Wiley*.1987.

Cole JC. How to deal with missing data. In Best practices in quantitative methods, *Journal Wiley Osborne (Ed.). Thousand Oaks, CA :Sage*, pp. 214–238. 2008.

Little RJA, Rubin DB. The analysis of social science data with missing values. *Sociological Methods and Research* ; **18** : pp.292–326. 1989.

Little R.J.A, Rubin D.B. Statistical Analysis with Missing. *A John Wiley & Sons. INC., Publication*. 2002.

Rubin L. H, Witkiewitz K, Andre J.S, Reilly S. Methods for handling missing data in the behavioral neurosciences : Don't throw the baby rat out with the bath water. *Journal of Undergraduate Neuroscience Education* ; **5**(2) : pp. A71-A77. 2007.

Schafer JL, Graham JW. Missing data : Our view of the state of the art. *Psychological Methods* ;**7**. 2002.

Dempster A, Laird N, Rubin D.B. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society Series B* ;**39**. 1977.

Rubin D. B. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. *Educational Testing Service*. 1978.

Schafer J. L. Analysis of incomplete multivariate data. *London : Chapman & Hall/CRC Press*. 1997.

Rubin D.B. Multiple imputation after 18+ years (with discussion). *Journal of American Statistical Association* ; **91** (432) : pp.473-489. 1996.

Schafer J. L. Multiple imputation : A primer. *Statistical Methods in Medecine Research* ; **8** : pp.3-15. 1999.

Zio Di M and Guarnera U. Semiparametric predictive mean matching. *AStA Advances in Statistical Analysis* ; **93**(2) : pp.175–186. 2009.

Rogier C, Tall A,Diagne N , Fontenille D, Spiegel A , Trape J.F. Plasmodium falciparum clinical malaria : lessons from longitudinal studies in Senegal. *Parassitologia* ;**41** (1-3) : 255-9. 1999.

Roussilhon C, Oeuvray C, Muller-Graf C, Tall A, Rogier C, Trape J F, Theisen M, Balde A Toure, Perignon J L, Druilhe P. Long-term clinical protection from falciparum malaria is strongly associated with igg3 antibodies to merozoite surface protein 3. *PLoS Medecine* ; **4**(11) : pp. e320, DOI : 10.1371/journal.pmed.0040320. 2007.

Toure B.A, Perlaza B-L, Sauzet J P *et al.* Evidence for multiple b- and t-cell epitopes inplasmodium falciparum liver-stage antigen 3. *Infection and Imminity* ; **77**(3) : pp. 1189-1196.2009.

Melanie G.C., Berchtold, A.. Imputation des donnes manquantes : comparaison de diffrentes approches. *Inria*. 2010.

Appendix : Tables and Figures

Tables

Id Subjet	PA	F15	MSP1	AMA1	MSP3	GLURP
AI-022/01	1.01	2.38	0.16	0.45	2.04	1.19
AI-022/02	2.44	3.66	NA	NA	0.97	2.85
AI-012/06	1.91	NA	NA	NA	NA	4.52
AI-006/09	2	5.33	0.17	NA	1.02	2.25
AI-007/03	1.6	2.33	0.14	0.42	NA	NA
AI-015/03	NA	1.83	0.14	0.42	NA	NA
AI-001/01	4.48	NA	0.36	2.1	2.25	1.8
AI-002/02	NA	1.55	NA	1.08	1.23	1.13
AI-003/05	NA	2.07	0.21	5.02	1.63	NA
AI-003/06	NA	2.51	0.21	0.8	1.38	1.45
AI-006/05	NA	1.84	0.16	NA	NA	0.94
AI-006/07	1.44	2.14	0.34	NA	1.2	0.76
AI-006/08	1.34	2.54	0.21	0.64	0.88	1.2

TABLE 6.1. Notations of missing values

	Missing data methods											
Missing	CC		Mean		Knn		MI Reg		MI EM		MI pmm	
data (%)	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
5	0.04	0.07	0.01	−0.03	−0.02	−0.04	0.01	0.02	0.05	−0.10	0.01	−0.03
10	−0.11	−0.16	−0.01	−0.05	−0.03	−0.12	0.01	−0.005	0.001	−0.003	−0.01	−0.05
15	−0.14	−0.08	−0.08	−0.19	−0.05	−0.09	−0.04	0.04	0.01	−0.003	−0.08	−0.19
20	−0.11	−0.08	−0.02	−0.10	−0.09	−0.13	−0.04	0.04	−0.02	−0.04	−0.02	−0.10
25	−0.12	−0.12	−0.01	−0.24	−0.07	−0.15	0.06	0.01	0.03	−0.07	−0.01	−0.24
30	−0.08	−0.03	0.01	−0.25	−0.11	−0.17	−0.01	−0.03	−0.02	−0.14	0.01	−0.25
35	−0.20	0.24	0.12	−0.23	−0.09	−0.18	0.09	0.1	0.05	−0.04	0.12	−0.23
40	−0.14	−0.06	−0.03	−0.32	−0.07	−0.18	−0.20	0.24	0.10	−0.04	−0.03	−0.32
45	−0.65	−0.51	−0.34	−0.37	−0.12	−0.20	−0.14	−0.06	0.06	−0.07	−0.34	−0.37
50	−0.12	−0.12	−0.01	−0.24	−0.12	−0.22	−0.65	−0.51	0.05	−0.07	0.01	−0.24

TABLE 6.2. Residual means and variances for missing data method and for each percent of missing data. Residual means and standard deviation were calculated by subtracting the estimate mean from the observed mean and by subtracting the estimate standard deviation from the observed standard deviation

.

	Missing data methods					
Missing	CC	I.Mean	I.Knn	MI.pmm	MI.EM	MI.Reg
data (%)	p.value	p.value	p.value	p.value	p.value	p.value
5	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
10	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
15	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
20	$1.02e-11$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
25	$1.49e-08$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
30	$1.37e-07$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
35	$2.33e-07$	$6.8e-16$	$3.8e-16$	$6.8e-16$	$< 2.2e-16$	$< 2.2e-16$
40	$3.8e-05$	$2.08e-13$	$2.08e-16$	$2.08e-13$	$< 2.2e-16$	$< 2.2e-16$
45	$5.0e-04$	$9.11e-10$	$9.11e-16$	$9.11e-10$	$< 2.2e-16$	$< 2.2e-16$
50	$1.37e-07$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$

TABLE 6.3. P-values for each missing technique and percent of missingness. The p-value of the complete dataset is inferior or equal to 2.210^{-16}

.

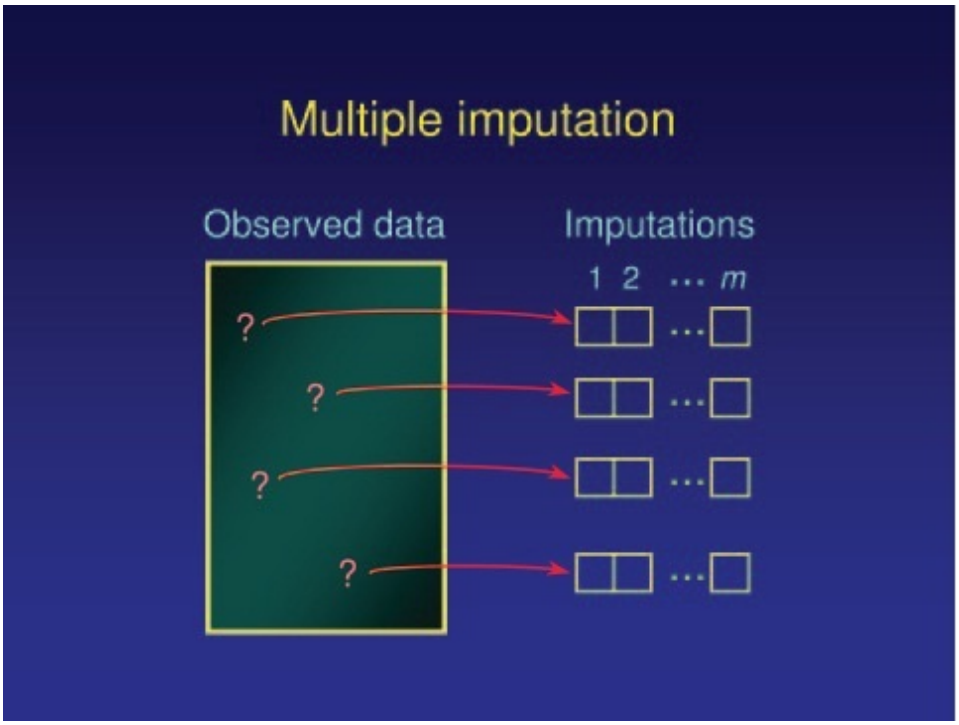

FIGURE 6.1. Scheme of multiple imputation, where question marks indicated missing data (Yulei (2010))

Figures

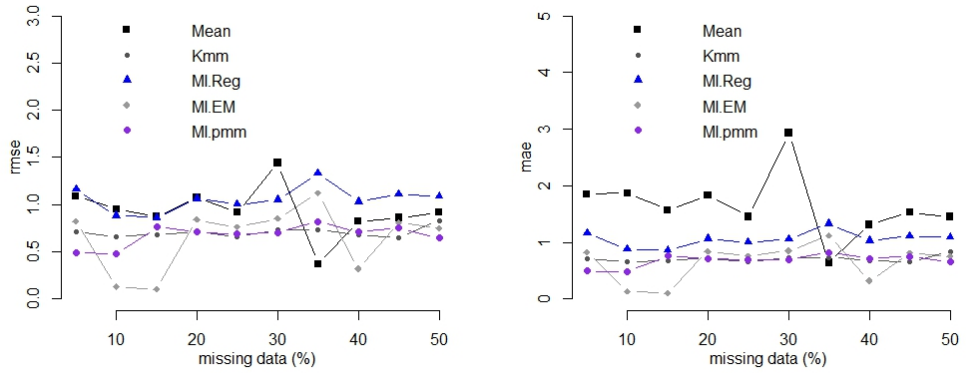

FIGURE 6.2. Root Mean Square errors and Mean-Square errors plotted as a function of percent of missing data and missing data methods.

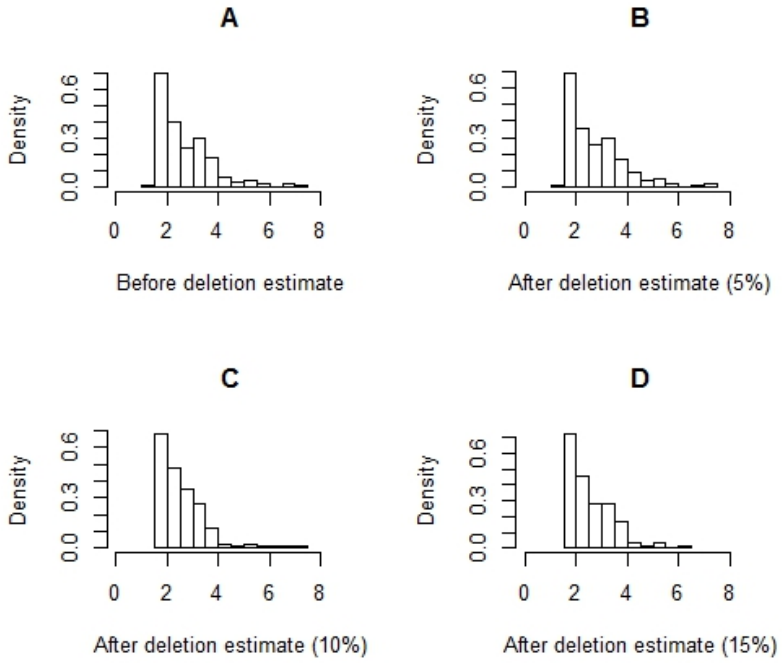

FIGURE 6.3. Histogram of linear regression coefficient before and after deleting some percent of missing data using listwise deletion

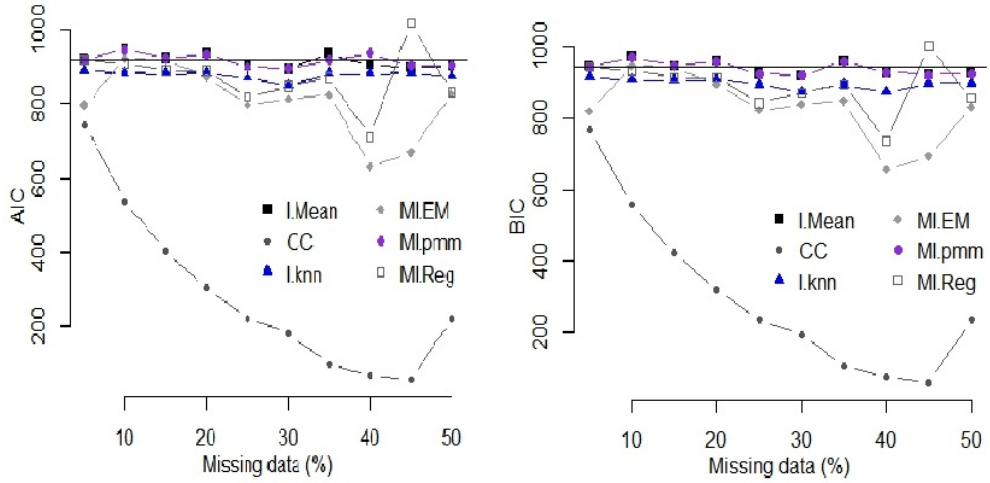

FIGURE 6.4. AIC and BIC of multiple regression model as a function of percent of missing data. Black horizontal lines mark AIC and BIC for the complete dataset

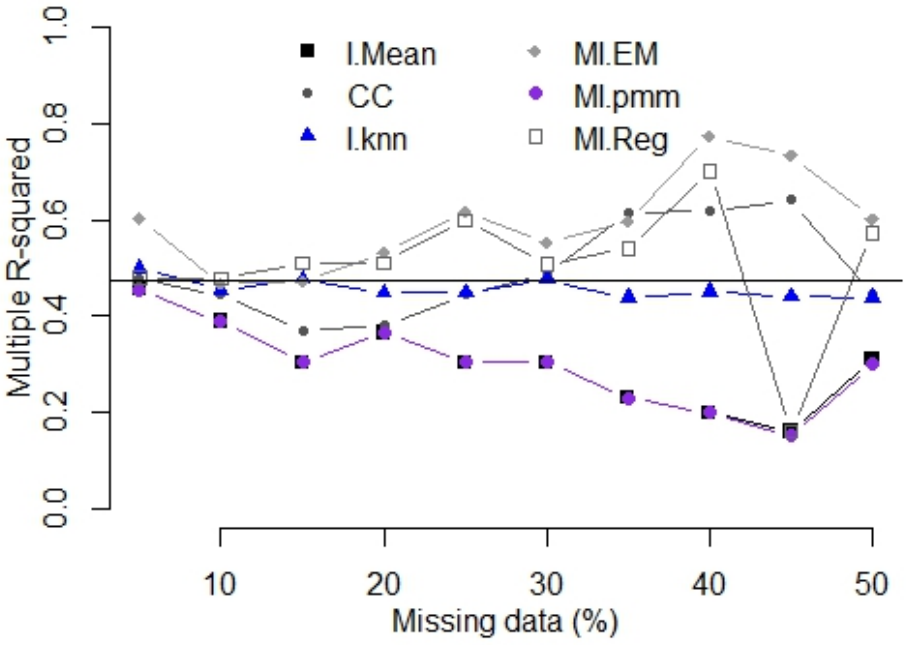

FIGURE 6.5. Multiple R-square as a function of percentages of missing data. Black horizontal line marks multiple R-square for the complete dataset