Gestion des données manquantes dans les études séro-épidémiologiques

Oumy Niass[1,2], Aïssatou Touré[1], Abdou Kâ Diongue[2], Aly Souleymane Dabye[2].

niass_oumy@yahoo.fr / oniass@pasteur.sn

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

I. Introduction

L'appréhension des données manquantes est un problème délicat. Non pas à cause de sa gestion informatique mais plutôt à cause des conséquences de leur traitement (suppression des individus ayant une mesure manquante ; ou remplacement par une valeur plausible à partir des observations disponibles : On parle d'imputation) sur les résultats d'analyse ou sur les paramètres d'intérêt. Les données manquantes peuvent se retrouver dans les variables à expliquer ou les variables indépendantes. Il existe plusieurs solutions aux problèmes des données manquantes. La méthode d'élimination est le mode de gestion le plus couramment utilisée (c'est la méthode par défaut de tous les logiciels statistiques usuels). Cette technique est raisonnable pour une proportion de données manquantes au moins égale à 5%, sinon elle peut, induire à une perte d'information ou introduire des biais dans les conclusions tirées des résultats d'analyse ou même empêcher la convergence statistique du modèle souhaité [1,2].
De nos jours, il existe de nombreuses méthodes alternatives de traitement de données manquantes, en passant par l'analyse cas-complet, les méthodes d'imputation dites simples et l'imputation multiple [3].
L'objectif principal de cette étude est de comparer les estimations de différentes méthodes d'imputation pour contourner le problème des données manquantes. Pour ce faire nous utilisons une base de données réelle dans laquelle nous allons appliquer les méthodes dites simples qui consistent à remplacer une valeur manquante par une seule valeur plausible et les méthodes d'imputation multiple.

II. Méthodes de traitement de données manquantes

Dans la littérature, il existe trois hypothèses distinctes sur l'origine du mécanisme de non réponse [3,4]; MCAR (Missing Completely At Random, si la probabilité de non réponse pour une variable donnée ne dépend pas de celle-ci, mais uniquement des paramètres extérieurs indépendants de cette variable ; MAR (Missing At Random), si la probabilité de non-réponse peut dépendre des observations mais pas des DM (Données Manquantes); et MNAR(Missing Not At Random), lorsque la probabilité de non-réponse est liée aux valeurs prises par la variable ayant des DM.

A. Méthodes d'imputation simples

1. Imputation par la moyenne ou médiane (I.Mean)

On attribut à chaque valeur manquante la moyenne ou la valeur médiane de la même variable. L'inconvénient de cette méthode est qu'elle conduit à une réduction systématique de la dispersion de chacune des variables et risque de briser d'éventuelles relations multidimensionnelles sous-jacentes entre les variables.

2. Imputation utilisant la régression (I.Reg)

Le principe de cette méthode est d'utiliser les observations renseignées pour créer un modèle de régression et ensuite utiliser le modèle pour prédire les données manquantes. Les variables avec données manquantes sont considérées comme des variables dépendantes. Les valeurs manquantes seront remplacées par les valeurs prédites selon le modèle.

Explicitement, soient $Y = (Y_{Obs}, Y_{mis})$ une variable avec des valeurs manquantes (Y_{mis}) et un modèle de régression linéaire simple basé sur les valeurs observées : $Y_{obs} = X\beta + \mu$ où $\mu \sim N(0, \sigma)$ avec

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, $X = (X_1, X_2, \ldots, X_p)$ et $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$

X représente la matrice des covariables (On suppose que X est complète) et β le paramètre d'intérêt.

Nous avons β^* et σ^* les estimations respectives de β et σ obtenues à partir du modèles sur les données observées. Nous générons les $\mu^* \sim N(0, \sigma^*)$ et nous prédisons les valeurs manquantes par :

$$Y_{mis} = X\beta^* + \mu^*$$

3. Imputation par la méthode du plus proche voisin (I.KNN)

On attribut à l'individu qui a une mesure manquante sur une variable donnée, la valeur de celui qui est le plus proche qui présente une mesure pour cette même variable. Cette similarité est habituellement définie par une fonction de distance entre les variables.

Soient Y notre variable d'intérêt, X la matrice des covariables et j, l'identifiant de l'individu n'ayant pas une mesure pour la variable Y (Y_j est manquant). On cherche parmi tous les individus i ayant une mesure à l'ensemble des variables l'individu j_0 qui minimise une certaine distance entre l'individu j et i :

$$j_0 = \text{Arg min}_{1 \leq i \leq P}\{d(i,j)\}$$

d est une mesure de distance, par exemple la distance euclidienne $d(i,j) = \sqrt{\sum_{k=1}^{p}(X_i^k - X_j^k)^2}$. Une fois que j_0 est déterminé, la valeur Y_{j_0} est attribuée à Y_j.

$$Y_j^* = Y_{j_0}$$

4. Imputation par l'algorithme EM (Espérance-Maximisation)

Initialement développé par A. P. Dempster, N.M. Laird et D. Rubin [5], L'algorithme EM est un algorithme itératif de calcul d'estimateur de vraisemblance par des modèles paramétriques lorsque les données sont

observées. Dans le cadre du traitement des données manquantes, Il permet de compléter les valeurs manquantes en se basant sur la vraisemblance maximale (maximum-likelihood estimation) de l'ensemble des données disponibles.

L'algorithme EM est une succession d'une étape d'espérance (E) où on évalue l'espérance de la log-vraisemblance pour la valeur courante du paramètre puis faire une actualisation du paramètre en celui qui maximise cette nouvelle fonction du paramètre : étape (M). L'estimation ainsi obtenues est celle qui maximise la probabilité d'observer ce qui a été réellement observée [6]. L'algorithme converge vers un point stationnaire sous des hypothèses de régularité.

B. Méthode par Imputation multiple

Elle a été proposée par Rubin en 1978 puis développée et décrite en détail par Rubin en 1987[3] et Schafer en 1997[7]. Elle consiste à remplacer une valeur manquante par m (m> 1) valeurs plausibles au sens d'un modèle statistique.

Rubin décrit la méthode comme une succession de 3 étapes. D'abord on attribue des valeurs aux données manquantes en utilisant un modèle aléatoire adapté. Ensuite répéter m fois cette étape afin d'obtenir les m tableaux de données complétées. Et enfin analyser ces m tableaux en utilisant les méthodes statistiques standards pour l'analyse des données complètes.

$$\beta_i^* = \frac{1}{m}\sum_{j=1}^{m} \beta_{i,j}^*$$

Plus le nombre m d'imputation est grand, plus les estimations seront précises. Cependant, Rubin (1987, 1996) a montré qu'en pratique à partir d'un faible nombre d'imputations (par exemple m=5) on a de bons résultats.

Dans ce travail nous allons appliquer cette méthode en utilisant :

- un algorithme basé sur le bootstrap, approchant des résultats de l'algorithme EM (IM.EM)
- l'approche "Predictive Mean Matching" (IM.pmm) [8]

III. Méthodologie

Dans le but de comparer les différentes méthodes d'imputation, nous disposons, une base contenant des données immunologiques du paludisme. Ces données vont constituer notre matrice de référence (MR) où il n'y a aucune valeur manquante. Nous avons choisit de manière aléatoire sur les n observations de la matrice MR des représentants de DM (données manquantes) au taux de T variant entre 0.5% et 50%. Ainsi nous avons créé des matrices avec valeurs manquantes (MVM). Pour chaque MVM les valeurs manquante simulées sont ré-estimées par la méthode de remplacement. La matrice MVM est ainsi complétée et nommée ME (Matrice Estimée). Enfin à chaque position des valeurs manquantes, Nous avons calculé la différence entre la valeur réelle et la valeur estimée.

Pour évaluer l'estimation, la moyenne des carrés d'erreur notée RMSE (Root Mean Square Error), l'erreur absolue moyenne, les moyennes et écart-types sont calculées pour chaque pourcentage de données manquantes.

$$RMSE_t = \sqrt{\frac{\sum_{i=1}^{M}(R_i - E_i)^2}{M}} \qquad MAE_t = \sqrt{\frac{\sum_{i=1}^{M}(R_i - E_i)}{M}}$$

Avec R_i la valeur réelle observée à la position où une valeur manquante a été insérée, E_i la valeur estimée par la méthode à cette même position et M correspond au nombre de valeurs manquantes dans la matrice MVM.

IV. Résultats d'analyses

1. Données

Nos données sont issues d'un suivi longitudinal fait sur 1448 enfants de moins de 10ans vivant dans huit villages (Aïdara, Daga Ndoup, Keur Ndianko, Keur saloly Bouya, Keur Samba Gueye, Némé Nding et Touba Nding) dans l'arrondissement de Toubacouta, dans la région de Fatick, dans le cadre du projet d'EDCTP. Sur un échantillon complet de 300 enfants, nous avons créé aléatoirement 10 bases incomplètes de taux de valeurs manquantes variant entre 5% et 50% et 290 bases complétées.
L'environnement de travail a été le logiciel R version 2.15.1.

2. Evaluation et comparaison de méthodes

L'ensemble des méthodes produisent des estimations avec un taux d'erreur moyenne assez faible. Ce taux est plus moindre avec la méthode d'imputation multiple par (pmm, EM) et celle des plus proches voisins.
Le taux d'erreur varie faible avec le pourcentage de valeur manquante pour les méthodes IM.pmm et I.KNN. Pour les méthodes I.Mean et I.Reg les taux d'erreur sont plus ou moins élevés.

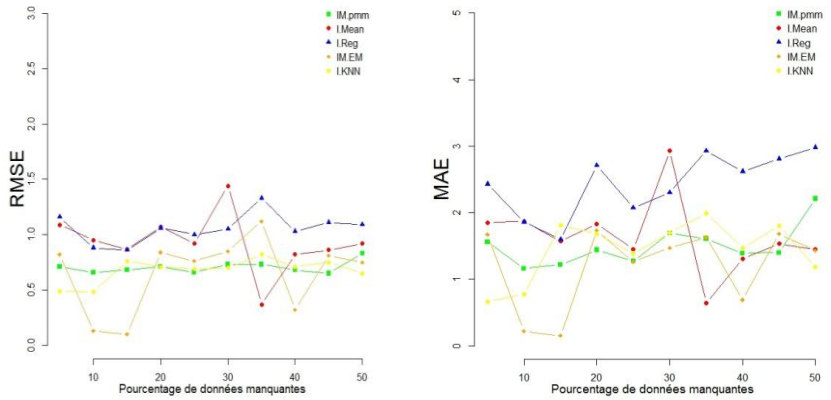

Fig1 : Evolution des RMSE et MAE en fonction du taux de données manquantes

Concernant l'estimation des moyennes, les méthodes I.Mean, I.Reg, ont des estimations plus centrées avec une légère sous-estimation. Mais en

résumé c'est les méthodes d'imputation multiple, pmm et EM et celle des plus proches voisins qui donnent des meilleurs résultats.

Pour l'estimation des écart-types (la dispersion), jusqu'à 20% de valeurs non observées, les méthodes IM.pmm, IM.EM et I.KNN donnent des estimations pratiquement sans biais. A partir de 25%, elles sous-estiment faiblement la variance.

V. Conclusion

Nos résultats montrent que les méthodes basées sur l'imputation multiple (IM .EM, IM.pmm) sont les meilleurs dans l'ensemble. Ces résultats semblent être en conformité avec d'autres études publiées récemment [9] La méthode basée sur les k plus proches voisin donne aussi des résultats satisfaisants.

Certaines méthodes d'imputation semblent influencer l'adéquation des modèles multivariés [10]. Dans cette perspective, nous prévoyons d'étudier l'impact de ces imputations sur les estimations des différents paramètres de différents Modèles multivariés.

VI. Bibliographie

1. Molenberghs G, Kenward M. Missing data in clinical studies. Wiley series in Probabilityans Statistics. Chichester: 2007.
2. Vergouw D, Heymans MW, Peat GM, Kuijpers T, Croft PR, de Vet HC, et al. The search for stable prognostic models in multiple imputed data sets. BMC Med Res Methodol 2010; 10:81.
3. Little R.J.A., Rubin D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley.
4. Little RJA, Rubin DB. Statistical analysis with missing data. Wiley series in Probability and Statistics. 2nd ed. New York: Wiley, 2002.
5. Little, R. J. and Schenker, N. (1995) : *Missing Data*, in Arminger, G. ; Clogg, C. C. and Sobel, M. E. : *Handbook of Statistical Modeling for the Social and Behavioral Sciences* ; Plenum Press, New York and London, 1995, chapter 2, pp. 39-75.
6. Allison P. D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods Research, 28(3), 301–309.
7. Schafer JL (1997). Analysis of incomplete multivariate data. Monographs on Statistics and Applied Probability 72. Chapman & Hall. London: 1997.
8. Zio Di.M., Guarnera. U. Semiparametric predictive mean matching: An empirical evaluation.ASCC., 2009.
9. Héraud-Bousquet V. PHD. Traitement de données manquantes en épidémiologie : Application de l'imputation multiple à des données de surveillance et d'enquêtes.tel-00713926, version 1-3 Jul 2012
10. Molenberghs G, Kenward M. Missing data in clinical studies. Wiley series in Probabilityans Statistics. Chichester: 2007.