Imputation des données manquantes : Comparaison de différentes approches.

Mélanie Glasson-Cicognani & André Berchtold
Université de Lausanne, Institut de Mathématiques Appliquées
SSP, Anthropole, CH - 1015 Lausanne
Melanie.Glasson@unil.ch & Andre.Berchtold@unil.ch

Résumé : Les données manquantes constituent un problème majeur, puisque l'information à disposition est incomplète et donc moins fiable. Notre objectif est de comparer par le biais de simulations numériques différentes méthodes existantes pour le traitement des données manquantes. En partant d'un fichier sans aucune donnée manquante, nous avons créé neuf scénarios variant en fonction du nombre de données manquantes et de leur type. Mille ensembles de données ont été générés à partir de chaque scénario, puis les données manquantes ont été traitées selon différentes approches. Les moyennes, écarts-types et corrélations entre variables imputées ont été comparés avec le fichier original sans données manquantes. L'influence du traitement des données manquantes sur un modèle de régression a aussi été évaluée. Nos résultats montrent que les méthodes basées sur l'imputation multiple sont globalement les meilleures. D'autres méthodes, comme par exemple l'imputation simple par régression, permettent aussi l'obtention de résultats intéressants, mais seulement dans certaines situations particulières.

Abstract : Missing data are a major concern, since available information is incomplete, hence less reliable. Our goal is to compare through numerical simulations different methods for the processing of missing data. Starting from a dataset without missing data, we created nine scenarios differing in function of the type and the number of missing data. One thousand datasets were generated from each scenario, and different approaches were applied on missing data. The mean, the standard error, and correlations between imputed variables were compared with the original dataset. The influence of the missing data treatment on regression model was also investigated. Results show that methods based on multiple imputation are overall the best approach. Other methods such as simple imputation using regression are also interesting, but for particular situations only.

Mots-clés : Données manquantes, imputation, régression, simulation numérique.

1 Introduction

Les données manquantes (DM) ont de multiples causes. Il peut être impossible de contacter une personne sélectionnée pour faire partie d'une enquête (non-réponse totale) ou un répondant peut refuser de répondre à une ou plusieurs questions (non-réponse partielle). Une mauvaise saisie de l'information peut également générer des DM. Finalement,

des DM peuvent aussi être causées par l'existence de données aberrantes qui doivent être supprimées avant d'effectuer des analyses.

L'absence de certaines informations pose un problème important pour les analystes, puisque l'information est incomplète, donc moins fiable. Il est nécessaire de traiter correctement les DM avant d'effectuer des analyses statistiques. Notre objectif est ici de comparer par simulation les performances de différentes méthodes d'imputation.

2 Types de données manquantes

Il existe plusieurs types de données manquantes, la classification la plus couramment utilisée ayant été proposée par Little et Rubin (1987) : "Missing completely at random" (MCAR) (complètement aléatoire), "Missing at random" (MAR) (aléatoire), "Missing not at random" (MNAR) (non aléatoire). Les DM sont MCAR lorsque la probabilité de non-réponse pour une variable ne dépend pas de celle-ci, mais uniquement de paramètres extérieurs, indépendants de cette variable. Cela veut dire qu'il n'est pas possible de définir un profil des individus ayant des DM et que la probabilité des DM est uniforme. De manière générale, ce type de DM est très rare. Les DM sont dites MAR lorsque la probabilité de non-réponse peut dépendre des observations mais pas des DM, par exemple s'il existe une différence de non-réponse entre les hommes et les femmes concernant la question du revenu, mais que parmi les hommes entre eux ou parmi les femmes entre-elles, la probabilité d'avoir des non-réponses est identique quel que soit le niveau du revenu. Finalement, les DM sont de type MNAR lorsque la probabilité de non-réponse est liée aux valeurs prises par la variable ayant des DM. C'est le cas par exemple lorsque les personnes ayant un revenu très élevé refusent beaucoup plus souvent de répondre à la question du revenu que les autres personnes.

3 Traitement des données manquantes

Face à la présence de données manquantes, la première possibilité consiste à exclure du fichier de données tous les individus ayant au moins une donnée manquante. Cela permet ensuite d'effectuer les analyses sur les cas dont toutes les données sont valides (analyse de cas complets). Une autre solution est l'imputation simple qui consiste à remplacer chaque donnée manquante par une valeur plausible. Par exemple, remplacer toutes les DM par la moyenne calculée sur les données réellement observées. D'autres méthodes d'imputation simple sont également disponibles, comme l'imputation par le plus proche voisin qui remplace les données manquantes par des valeurs provenant d'individus similaires pour lesquels toute l'information a été observée, et l'imputation par régression qui consiste à remplacer les DM par des valeurs prédites selon un modèle de régression. Il existe cependant de sérieuses contre-indications à l'application de certaines de ces méthodes (Schafer

& Graham, 2002). Certaines de ces méthodes ont été améliorées en ajoutant une marge d'erreur aléatoire, afin que l'imputation reflète mieux l'incertitude liée aux DM.

Durant les dernières décennies, de nouvelles méthodes d'imputation plus performantes ont été développées, en particulier l'imputation multiple (IM ; Rubin, 1987), dont le principe est de procéder à $m > 1$ imputations afin obtenir m valeurs pour chaque donnée manquante, et à combiner ensuite les statistiques calculées indépendamment sur les m jeux de données. L'utilisation correcte de l'IM peut demander un investissement important de la part de l'utilisateur, afin de s'assurer du respect de ses conditions d'application (Donzé, 2001). Par ailleurs, les résultats peuvent varier selon les logiciels et les modèles utilisés (Allison, 2000), mais cette méthode se révèle souvent meilleure que l'imputation simple. L'imputation multiple peut être réalisée sur la base de plusieurs modèles, comme par exemple une régression linéaire. L'algorithme EM est aussi couramment employé pour l'estimation par le maximum de vraisemblance de données incomplètes, tant en imputation simple qu'en imputation multiple, tout comme l'algorithme MCMC (Allison, 2003). Finalement, le "Predictive Mean Matching" (Di Zio & Guarnera, 2008) et l'"Approximate Bayesian Bootstrap" (Donzé, 2001) allient des approches paramétriques et non-paramétriques.

4 Les données

Nous utilisons les données "Boston Housing Data" disponibles librement à l'adresse http ://archive.ics.uci.edu/ml. Elles concernent la valeur de l'immobilier et d'autres caractéristiques de la périphérie de Boston. Ces données comportent 506 observations sur 14 variables. Il n'y a pas de données manquantes dans le fichier original. Nous avons choisi trois variables afin de générer artificiellement des DM, les variables V1 (CRIM), V3 (INDUS) et V7 (AGE). Avant de générer les DM, nous avons standardisé toutes les variables afin de permettre une comparaison plus aisée au niveau des résultats.

Neuf groupes de jeux de données incomplets ont été créés en combinant à la fois les types de DM et leur nombre. Les jeux de données sont d'abord différenciés au niveau du type de DM :

- Dans le premier cas, MCAR, nous avons créé les DM pseudo-aléatoirement, sans autre condition.
- Pour les jeux de données MAR, nous avons créé des DM pour chaque variable selon des conditions plus précises : les données manquantes de V1 dépendent de V11 (PTRATIO), celles de V3 et V7 dépendent de la variable V8 (DIS).
- Dans le cas MNAR, la présence de DM est conditionnée par les variables elles-mêmes.

Nous avons également fait varier le nombre de DM pour obtenir à nouveau 3 situations différentes : lorsque chaque variable contient 25 DM sur les 506 observations, 50 DM et finalement 100 DM. Pour chacune des 9 situations possibles, 1000 jeux de données ont été générés.

5 Méthodes

Nous avons choisi 9 méthodes de traitement des données manquantes largement répandues à l'heure actuelle, y compris des méthodes connues pour être peu performantes mais cependant toujours utilisées. Tous les calculs ont été effectués à l'aide du logiciel libre R (http ://www.r-project.org/).

1. Analyse des cas complets (CC) : La fonction n'est pas issue d'une librairie particulière.
2. Imputation par la moyenne (MEAN) : La fonction n'est pas issue d'une librairie particulière.
3. Imputation par la médiane (MED) : La fonction n'est pas issue d'une librairie particulière.
4. Imputation par régression simple (REG) : La fonction n'est pas issue d'une librairie particulière.
5. Imputation multiple par Markov Chain Monte-Carlo (MCMC) : La fonction d'imputation par MCMC est intégrée à la librairie R nommée *sbgcop*.
6. Imputation par le plus proche voisin (KNN) : La librairie R utilisée est *yaImpute*.
7. Imputation par le plus proche voisin "randomForest" (KNNRF) : Cette méthode est proche de la précédente, mais elle fait appelle à la libraire "randomForest" pour la définition de la distance entre observations.
8. Imputation multiple par un algorithme basé sur le bootstrap, approchant des résultats de l'algorithme EM (EM) : Cette méthode est réalisée grâce à la librairie *Amelia* avec $m = 10$.
9. Imputation multiple par "Predictive Mean Matching" (PMM) : Nous avons utilisé l'implémentation disponible dans la librairie R *mice* avec $m = 10$.

Dans un premier temps, nous avons comparé les caractéristiques des variables sur lesquelles des données ont été imputées avec les caractéristiques des variables originales correspondantes. Ensuite, nous nous sommes intéressés à l'influence des données manquantes sur les résultats d'un modèle de régression linéaire. A titre de point de référence, nous avons commencé par estimer le modèle de régression sur le fichier original sans DM. Nous avons ensuite réestimé le même modèle à partir de chacun de jeux de données dans lesquels des DM ont été créées, puis imputées. Nous avons comparé les résultats obtenus sur les fichiers traités par chaque méthode selon la précision de la distribution des statistiques et estimateurs par rapport à la référence issue des données originales.

6 Résultats

Nous nous concentrons ici sur les résultats obtenus lorsqu'il y a 100 données manquantes par variable. En ce qui concerne les caractéristiques des variables, nous résumons

nos résultats comme suit :

- Concernant l'estimation de la moyenne des variables ayant des données manquantes, la méthode MEAN produit des estimations centrées sur la référence uniquement en cas MCAR, sinon elle sur- ou sous-estime la valeur de la moyenne. La méthode MED a le même défaut et son point fort réside en MAR. De manière générale, les méthodes MEAN, MED et CC sont moins intéressantes, car leurs résultats sont plus dispersés et/ou biaisés. La méthode KNNRF offre des résultats plus stables que les autres méthodes en cas MNAR. Pour résumer, les méthodes PMM, REG, EM et KNNRF sont les plus intéressantes, en particulier, KNNRF en situation MNAR.
- Pour l'estimation de la dispersion des variables par l'écart-type, les méthodes PMM et EM sont généralement les plus efficaces, ainsi que REG et MCMC en cas MNAR. Au niveau de la précision et de la dispersion des estimations, c'est la méthode PMM qui est la plus fiable. La méthode REG, intéressante en cas de MNAR, a par contre le défaut, contrairement à PMM, de plus facilement sous-estimer la variance. Pour résumer, nous considérons que la méthode PMM est la plus intéressante pour l'estimation de l'écart-type de nos variables.
- Concernant l'estimation des corrélations entre les variables, nous constatons que les méthodes les plus précises en moyenne sont PMM et EM. Cependant, selon le type de DM, des méthodes comme REG et MCMC offrent parfois de bons voir de meilleurs résultats. Globalement, nous retenons la méthode PMM qui offre une bonne stabilité tant au niveau du nombre de DM qu'au niveau du type de DM.

Les principaux résultats relatifs au modèle de régression sont les suivants :

- Le R^2 est généralement bien estimé. Seule CC se distingue des autres méthodes tant au niveau de la précision (moyenne et médiane parfois inférieure à la référence) qu'au niveau de la dispersion nettement supérieure d'un jeu de données à l'autre.
- Les coefficients de régression sont en général mieux estimés par les méthodes PMM, MCMC et MED et les distinctions entre les méthodes sont déjà moins nettes en cas de MCAR ou MAR. Au niveau de la précision, les méthodes REG, MED et dans une moindre mesure MCMC sont moins stables que PMM et EM, car elles sous-estiment plus fréquemment la valeur du coefficient. De manière générale, nous concluons que les méthodes d'imputation multiple sont les plus fiables, quel que soit le type de DM.
- Concernant les p-valeurs des tests de significativité des coefficients de régression, nous constatons qu'elles sont quasiment systématiquement sous-estimées pour les variables à DM. Les méthodes les plus intéressantes sont REG, EM, MCMC et PMM. Les meilleures performances ont systématiquement lieu en cas de MAR et les pires en MNAR. La méthode d'analyse CC n'est vraiment pas intéressante, car elle présente l'inconvénient de produire des résultats beaucoup plus dispersés que toutes les autres méthodes, ceci en particulier pour les p-valeurs des variables sans

DM. En résumé, nous retenons ici les méthodes PMM, MCMC, EM et REG.

7 Conclusion

Nos résultats confirment le problème principal qui se pose lors du traitement des données manquantes : lorsque les données sont MNAR, la marge d'erreur est plus importante. Il arrive facilement que toutes les estimations soient supérieures ou inférieures à la valeur de référence. Cela nous pousse à la plus grande prudence quand aux conclusions tirées des résultats obtenus sur des données incomplètes.

Les méthodes qui permettent globalement d'arriver aux résultats les plus satisfaisants sont l'imputation multiple par PMM et MCMC. Les méthodes comme REG ou KNNRF sont intéressantes ponctuellement, pour l'estimation de l'écart-type des variables par exemple, car elles peuvent parfois permettre d'obtenir de bons résultats, notamment dans le cas MNAR.

Il est intéressant de constater qu'une méthode comme l'imputation simple par régression reste assez performante. Etant donné que PMM, basée sur la régression également, donne de bons résultats, il est probable que ces méthodes soient particulièrement bien adaptées pour des données corrélées et que cela les favorise plus que nécessaire. De plus, il est intéressant d'imputer les DM sur la base de modèles similaires à celui utilisé lors de l'analyse (Rubin, 1987 ; Allison, 2000). A l'inverse, les moins bons résultats pour la régression sont systématiquement obtenus avec les méthodes CC, KNN et KNNRF.

Bibliographie

[1] Allison P. D. (2000). Multiple Imputation for Missing Data : A Cautionary Tale. *Sociological Methods Research*, 28(3), 301–309.

[2] Allison P.D. (2003). Missing Data Techniques for Structural Equation Modeling. *Journal of Abnormal Psychology*, 112(4), 545–557.

[3] Di Zio M., Guarnera U. (2009). Semiparametric predictive mean matching. *AStA Advances in Statistical Analysis*, 93(2), 175–186.

[4] Donzé L. (2001). L'imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse des données : l'enquête 1999 KOF/ETHZ sur l'innovation. Ecole polytechnique fédérale de Zurich, Centre de recherches conjoncturelles.

[5] Schafer J.L., Graham J.W. (2002). Missing Data : Our View of the State of the Art. *Psychological Methods.*, 7(2), 147–177.

[6] Little R.J.A., Rubin D.B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley.

[7] Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley.