Analyse de données manquantes dans un contexte longitudinal

UNIL | Université de Lausanne
Faculté des sciences sociales et politiques

Mélanie Cicognani - melanie.cicognani@unil.ch

Sous la direction d'André Berchtold (IMA) & Jean-Marie Le Goff (Labo. PAVIE-UniL)

Institut de Mathématiques Appliquées, Faculté des SSP, Université de Lausanne

Introduction

Les données d'enquête en sciences sociales sont souvent incomplètes : certaines réponses manquent ou ne sont pas valides ; ce sont des données manquantes. Ces données ne peuvent pas être analysées de la même manière que des données complètes. Le problème des données manquantes est récurent. Il influe directement sur la représentativité des échantillons et introduit un biais dans les analyses, les résultats et les conclusions fondés sur ces données. Il est nécessaire de traiter ces données manquantes. **Nous voulons étudier les méthodes de traitement de ce problème et proposer des solutions performantes, notamment dans le cas longitudinal.**

Motivations

Notre travail est motivé par l'enquête "devenir parent" menée par le centre lémanique d'étude des parcours et modes de vie (PAVIE). Durant cette enquête, des couples devaient répondre à des questions à trois reprises (vagues 1, 2 et 3). Selon les informations disponibles actuellement (voir Figure 1), sur 236 couples ayant participé en vague 1, seulement 168 ont répondu à l'appel en vague 2 et environ 174 lors des derniers entretiens. Afin d'améliorer l'analyse de ces données longitudinales, il est souhaitable de pouvoir reconstituer certaines des données manquantes de la vague 2.

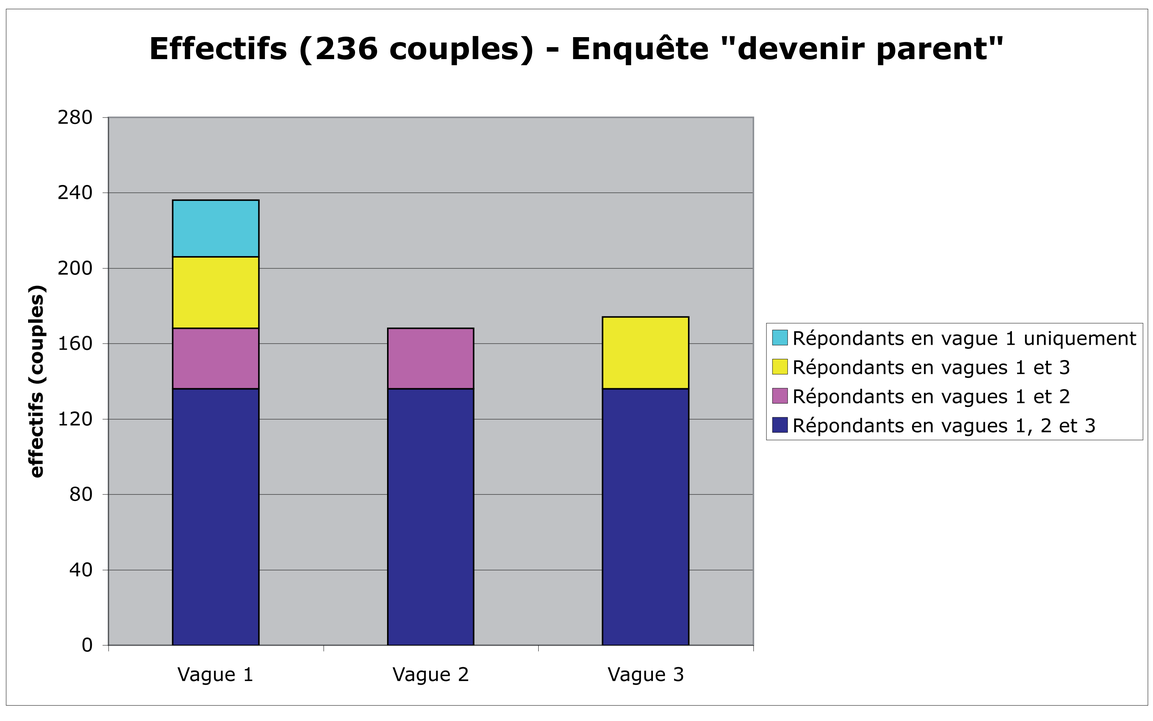

FIGURE 1: Nombre de couples participant à l'enquête "devenir parents" en fonction de la vague.

Méthode

Nous avons commencé par travailler sur un modèle simple de régression linéaire, afin de comparer des méthodes de traitement de données manquantes couramment utilisées. Nous sommes partis de données déjà existantes, sans données manquantes, puis nous avons créé 1000 nouveaux fichiers en générant plusieurs types de données manquantes. Nous étudions de manière similaire différents modèles comportant des variables ordinales ainsi que des données longitudinales.

Résultats

Nos premiers résultats sont conformes à nos attentes. En ayant 10 % de données manquantes, il est en effet fortement déconseillé de ne rien faire et d'effectuer les analyses en ne tenant compte que des individus ayant répondu à toutes les questions. Par ailleurs, remplacer les données manquantes par la moyenne ou la médiane amène un biais considérable aux résultats.

Considérons par exemple la distribution de la moyenne d'une variable pour 1000 échantillons comportant des données manquantes de type MAR (au hasard). **Les résultats sont mauvais tant pour la méthode de remplacement par la médiane ou la moyenne que sans remplacement du tout, avec une nette surestimation de la vraie valeur moyenne. En revanche, les méthodes modernes basées sur l'imputation multiple donnent de bien meilleurs résultats.**

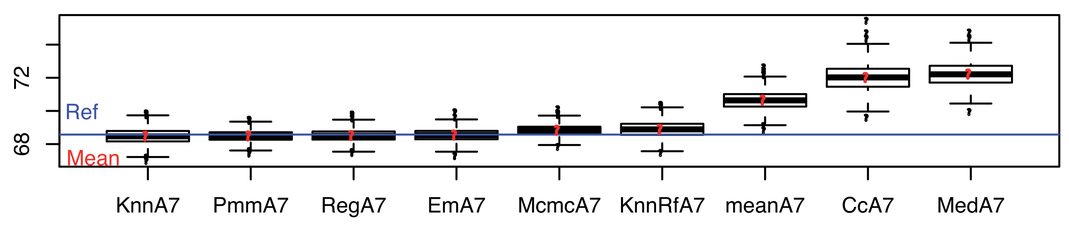

FIGURE 2: Moyenne de la variable calculée sur 1000 échantillons à l'aide de différentes méthodes. PmmA7 : imputation multiple + predictive mean matching ; EmA7 : imputation multiple + EM ; MeanA7 : imputation par la moyenne ; CcA7 : aucune imputation ; MedA7 : imputation par la médiane.