Chapitre 4

Données manquantes

4.1 Introduction

Dans la phase de préparation des données, parallèlement ou en amont de la sélection de variables, on doit considérer le problème du nettoyage de la base de données. Il s'agit d'identifier les données aberrantes, les individus atypiques [1] et de traiter (ou gérer) les données manquantes.

4.1.1 Données aberrantes

Les données abberantes peuvent prendre plusieurs formes.
- Données catégorielles
- Données positives
- Valeur extrème de probabilité très faible

Dans le cas de données issues d'une distribution continue multivariée, on peut repérer les données aberrantes par projection sur un sous espace (ACP, ACP non linéaire) ou éventuellement par classification (présence de classe d'effectif très faible). On doit alors considérer les données correspondantes comme des données manquantes.

Remarque : il est parfois difficile de savoir si une données est aberrante ou atypique.

4.1.2 Induvidus atypiques textitoutliers

En ce qui concerne la **détection d'individus atypiques**, le problème a été évoqué dans le cadre des travaux dirigés. Le solution proposée consiste à réaliser une analyse en composante principale non linéaire pour mettre en évidence ces individus. On peut aussi les retrouver dans une classe d'effectif très faible dans une classification. On peut alors soit les considérer comme des individus supplémentaires de façon à ne pas biaiser les analyses ou les estimations, soit transformer les données pour 'symétriser' sa distribution.

Exemple - Données démographiques à l'échelle de la commune : on divise les données mesurées en nombre d'individus par la population de la commune et celles mesurées en nombre de ménages par le

1. en anglais : outliers

nombre de ménages total de la commune. On obtient ainsi des données comprises entre 0 et 1. Les communes importantes peuvent encore avoir un poids (trop) important...

4.1.3 Données manquantes

En statistique, on parle de valeur manquante lorsqu'on n'a pas d'observations pour une variable donnée pour un individu donné. Le problème de la gestion des données manquantes est un vaste sujet. Les données manquantes ne peuvent pas être ignorées lors d'une analyse statistique. Mais selon leur proportion et leur type, des solutions différentes vont être choisies. On pourra soit retirer les variables ou les individus présentant des données manquantes ou imputer des valeurs aux données manquantes ou encore développer des méthodes (ou algorithme) qui permettent de mener les analyses en présence de données manquantes.

Différents types de données manquantes

- Doonées manquantes complètement aléatoirement [2] - Les données sont manquantes complètement aléatoirement si probabilité qu'une observation soit manquante ne dépend pas des mesures observées ou non observées. En termes mathematiques, ça s'écrit

$$P(r|\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{miss}}) = P(r)$$

  où r représente la réponse. On dit parfois que la non réponse est répartie uniformément. Utilisation de covariables : age + utilisation du prénom. Un exemple en météo (date+autre variable).
- Données manquantes aléatoirement [3] - Les donnés sont manquantes aléatoirement si, sachant données les données observées, le mécanisme de non réponse ne dépend pas des données non observées. Mathématiquement, on écrit

$$P(r|\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{miss}}) = P(r|\mathbf{x}_{\text{obs}})$$

  On remarque que dans ce cas, on peut mener des analyses en utilisant uniquement l'information observée. Exemples de données manquantes aléatoirement :
  - Un sujet peut être retiré d'un essai si sa condition n'est pas assez bien controlée (selon des critères prédéfinis sur la réponse).
  - Deux mesures de la même variable sont réalisées en même temps. Si elles diffèrent de plus d'un écart pré défini, une troisième est retenue. La troisième mesure est manquante dans le cas où les deux premières ne diffèrent pas (ou pas trop).

  Les méthodes basées sur la vraisemblance restent valides pour les jeux de données présentant des données manquantes aléatoirement. Cependant, les méthodes basées sur d'autres critères ne le sont pas. En particulier, l'estimation des moments et autres statistiques simples, va être biaisée.
- Données manquantes non aléatoirement [4] - Ce cas correspond à une mécanisme de non réponse non-ignorable. Ca signifit que
  1. Même en tenant compte de tout l'information observée, les raisons pour lesquelles des données sont manquantes dépendent des données manquantes elles mêmes. Exemple : un

2. en anlgais : Missing Completely At Random (MCAR)
3. en anglais : Missing At Random (MAR)
4. en anglais : Missing Not At Random (MNAR)

adolescent qui sort, de lui même, d'un essai longitudinal sur l'obésité parce qu'il constate qu'il a grossit.

2. Pour obtenir des estimations valides, un modèle joint des données complètes sachant le mécanisme de réponse est nécéssaire.

Malheureusement

1. On ne peut généralement pas dire, à partir des données, quel est le méchanisme de manque (MCAR, NMAR ou MAR).
2. Dans le cas MNAR, il est rare que l'on connaisse le modèle associé au manquement.

Les solutions quand on a des données manquantes (MCAR).

1. Obtenir les réponses (coûteux ou impossible).
2. Remplacer les données manquantes : imputation.
3. Utiliser/développer des méthodes adaptées.

4.2 Données manquantes et imputation

L'imputation regroupe les méthodes utilisées pour remplacer les données manquantes.

4.2.1 Imputation par la moyenne

Les méthodes d'imputation les plus simples consistent à remplacer les données manquantes par leur moyenne ou leur médiane. L'inconvénient de cette approche est qu'elle conduit à une sous-estimation parfois violente de la variance des estimateurs.

Exemple - On tire un échantillon de 10 observations suivant une loi de Gauss de moyenne nulle et d'écart-type 10. Puis on tire les indices correspondant aux valeurs manquantes selon une loi uniforme sur $\{1, \cdots, 10\}$. On choisit d'avoir 3 valeurs manquantes sur 10 soit environ 30% de valeurs manquantes.

Données complètes										Moyenne	Ecart-type
7	-5	21	-1	1	10	0	0	-8	2	2.7	8.2
Données incomplètes (ex 1)										Moyenne	Ecart-type
7	-5	21	-1	1	10	-	-	-8	-	3.3	8.1
Données incomplètes (ex 2)										Moyenne	Ecart-type
-	-5	21	-	1	10	0	0	-	2	4.3	6.9

En répétant 1000 fois cette opération, on estime l'écart-type de l'estimateur de la moyenne après imputation et on obtient une valeur de 1.59 alors que $\hat{\sigma}/sqrt10 = 2.60$. On sousestime très largement la variance de l'estimateur de la moyenne.

4.2.2 Imputation par tirage conditionnel

On peut améliorer l'idée de l'imputation par la moyenne en réalisant de l'imputation par tirage conditionnel. Le principe est d'utiliser l'information apportée par les variables renseignées.

Plusieurs approches sont possibles :

1. Estimer la loi jointe et générer conditionnellement une réalisation pseudo aléatoire de cette loi. Mais il est généralement difficile d'estimer une loi jointe au délà de 2 ou 3 variables. Une alternative intéressante et plus facile à mettre en oeuvre, bien qu'éventuellement couteuse, consiste à utiliser une méthodes des plus proches voisins.
   - Soit i l'individu présentant une non réponse.
   - Calculer la distance de i à tous les individus ayant les mêmes variables renseignées.
   - Retenir les k plus proches voisins.
   - Imputer la moyenne des k plus proches voisins à la donnée manquante.
2. Réaliser une classification à partir des *variables complètement renseignées* et estimer la moyenne conditionnelle par classe. On peut voir cette méthode comme une sorte généralisation de la méthode des plus proches voisins. Dans les deux cas, l'imputation va se baser sur les observations les plus proches.
3. Construire un modèle de régression à partir des *individus complètement renseignés* et l'utiliser pour prédire les données correspondant aux données manquantes.

De façon générale, il est préférable de faire de l'**imputation multiple**. L'idée est de réaliser plusieurs tirages et de répéter les analyses pour prendre en compte et retablir la variabilité sous jacente à l'absence de données. L'usage est de faire 5 tirages...

4.2.3 Imputation par analyse factorielle

Considérons le cas de données issues de variables quantitatives. L'analyse en composante factorielle permet de *'reconstruire'* des données par projection dans un espace de dimension réduite. Cette caractéristique peut-être exploiter pour remplacer des données manquantes.

L'approche la plus naïve consiste à estimer la matrice de covariance à partir des individus renseignés puis d'estimer les paramètres de l'analyse en composante principale et enfin à reconstruire les données manquantes.

Remarques sur l'estimation de la matrice de covariance

1. Chaque coefficient est estimé à partir de seulement 2 colonnes de la table de données on n'est donc pas amené à supprimer des colonnes ou lignes entières si elles sont au moins partiellement renseignées.
2. L'estimation reste de moins bonne qualité que si toutes les données étaient renseignées.