METHODES D'IMPUTATION DES DONNEES MANQUANTES

DIOP Ousseynou

Année 2022-2023

Section 1

Introduction

Introduction

Malgré la quantité croissante de données disponibles et l'émergence du BigData, les problématiques de données manquantes restent très répandues dans les problèmes statistiques et nécessitent une approche particulière. Ignorer les données manquantes peut entraîner, outre une perte de précision, de forts biais dans les modèles d'analyse,des erreurs dans les résultats de l'analyse pouvant conduire ainsi a des conclusions erronées.Pour résoudre ce problème, les méthodes d'imputation ont été développées pour estimer les valeurs manquantes en utilisant des informations disponibles dans la base.

Dans cette présentation , il sera question de definir les différents types de données manquantes et illustration de leurs répartitions possibles, de decrire les principales stratégies de gestion des données manquantes par suppression de données ou par d'autre méthode stochastique, sans souci d'exhaustivité .

Section 2

I.Definitions

I.1 Données manquantes

Les données manquantes sont des valeurs qui ne sont pas présentes dans un ensemble de données. Elles peuvent être manquantes pour différentes raisons, comme une erreur de saisie, un échec de mesure, ou un refus de réponse.

I.2 Typologie des données manquantes

Afin d'aborder correctement l'imputation des données manquantes il faut en distinguer les causes, surtout si elles ne sont pas le simple fruit du hasard.Il existe trois catégories de données manquantes:

I.2.1 MCAR (missing completely at random)

Une donnée est MCAR, c'est-à dire manquante de façon complètement aléatoire si la probabilité d'absence est la même pour toutes les observations. Cette probabilité ne dépend donc que de paramètres extérieurs indépendants de cette variable.

Par exemple : si chaque participant à un sondage décide de répondre à la question du revenu en lançant un dé et en refusant de répondre si la face 6 apparaît [1]. À noter que si la quantité de données MCAR n'est pas trop importante, ignorer les cas avec des données manquantes ne biaisera pas l'analyse. Une perte de précision dans les résultats est toutefois à prévoir

I.2.2 MAR (Missing at random):

Le cas des données MCAR est peu courant. Il arrive lorsque les données ne manquent pas de façon complètement aléatoire; si la probabilité d'absence est liée à une ou plusieurs autres variables observées, on parle de missingness at random (MAR). Il existe des méthodes statistiques appropriées qui permettrons d'éviter de biaiser l'analyse

I.2.3 MNAR (Missing not at random) :

La donnée est manquante de façon non aléatoire (MNAR) si la probabilité d'absence dépend de la variable en question. Un exemple répandu est le cas où des personnes avec un revenu important refusent de le dévoiler. Les données MNAR induisent une perte de précision (inhérente à tout cas de données manquantes) mais aussi un biais qui nécessite le recours à une analyse desensibilité.

I.3 Imputation

L'imputation est un processus qui consiste à remplacer les valeurs manquantes dans un ensemble de données par des valeurs estimées afin de maintenir l'intégrité des données et d'obtenir des résultats plus précis et plus fiables.

Section 3

II.Méthodes d'imputation

II.Méthodes d'imputation

Il existe plusieurs méthodes d'imputation de données dans RStudio. Les méthodes d'imputation les plus courantes sont:

- Imputation par suppression;
- Imputation par la moyenne;
- Imputation par la médiane;
- Imputation par régression;
- Imputation par la méthode LOCF;
- Imputation par la méthode « hot-deck »;
- Imputation par la méthode de Monte carlo;
- Imputation par la forêt aléatoire;
- imputation par la méthode du plus proche voisin ;
- Imputation par l'analyse en composantes principales (ACP).

II.1 Méthode d'imputation par suppression

L'imputation par suppression consiste à supprimer les individus avec des valeurs manquantes de l'analyse.

II.2 Méthode d'imputation par la moyenne

Cette méthode remplace les valeurs manquantes par la moyenne des valeurs existantes pour cette variable.

II.3 Méthode d'imputation par la médiane

L'imputation par la mediane est similaire à l'imputation par la moyenne, sauf qu'elle utilise la médiane des valeurs présentes dans l'ensemble de données au lieu de la moyenne.

II.4 Méthode d'imputation par la régression

Elle consiste à estimer les valeurs manquantes à l'aide d'un modèle de régression basé sur les autres variables de l'ensemble de données.

II.5 Méthode d'imputation par « last observation carried forward » (LOCF)

Cette méthode consiste à remplacer les valeurs manquantes par la dernière observation valide connue pour cette variable.

II.6 Imputation par la méthode de Monte carlo

La méthode de Monte Carlo consiste à simuler plusieurs valeurs possibles pour chaque valeur manquante, en utilisant une distribution probabiliste basée sur les valeurs présentes dans l'ensemble de données.

II.7 Imputation par la méthode de forêt aléatoire

L'imputation par la forêt aléatoire est une méthode d'imputation des données manquantes qui repose sur l'utilisation de l'algorithme de la forêt aléatoire (une forêt aléatoire (ou Random Forest en anglais) est un modèle d'apprentissage automatique de type ensembliste, c'est-à-dire qu'il combine plusieurs arbres de décision pour obtenir une meilleure performance de prédiction).

II.8 imputation par la méthode du plus proche voisin

Ici, il s'agit d'estimer les valeurs manquantes en utilisant les valeurs les plus proches dans l'ensemble de données, mesurées en termes de distance.

II.9 Imputation par l'analyse en composantes principales (ACP)

Imputation par ACP est le remplacement des valeurs manquantes en utilisant les coefficients de régression obtenus à partir de l'analyse en composantes principales de l'ensemble de données. L'ACP (Analyse en Composantes Principales) est une méthode d'analyse statistique qui permet de réduire la dimensionnalité d'un ensemble de données en transformant les variables corrélées en un ensemble de variables non corrélées appelées composantes principales.

Section 4

III. Bibliothèques

III. Bibliothèques

Il existe plusieurs bibliothèques R qui offrent des méthodes pour imputer les données manquantes.On peut citer les bibliothèques suivantes:

- Mice;
- Amelia;
- imputeTS;
- Hmisc;
- missForest;
- miceadds.
- zoo

Section 5

IV. Limites d'imputations des données manquantes

IV. Limites d'imputations des données manquantes

Bien que l'imputation de données manquantes soit une technique utile pour traiter les données incomplètes, elle présente également certaines limites.

- Les résultats peuvent être biaisés;
- les imputations peuvent être inexactes ;
- les méthodes d'imputation peuvent ne pas convenir à tous les types de données manquantes;
- l'imputation peut nécessiter beaucoup de temps et de ressources;
- les données imputées peuvent être difficiles à interpréter.
- l'imputation des données manquantes peut entraîner une perte d'information.
-

Section 6

Conclusion

Conclusion

Il est important de prendre en compte les données manquantes lors de l'analyse des données, car cela peut affecter les résultats et les conclusions tirées à partir des données. Des méthodes d'imputation appropriées peuvent être utilisées pour remplacer les valeurs manquantes et améliorer la qualité des données pour l'analyse.Ces méthodes d'imputation peuvent être basées sur des approches statistiques ou non statistiques et peuvent varier en fonction de la complexité des données, du niveau de précision souhaité et des objectifs de l'analyse.Il faut noter que l'imputation peut introduire des erreurs dans l'ensemble de données, en particulier si les méthodes utilisées ne sont pas adaptées ou si les données manquantes sont mal interprétées. Par conséquent, il est important de comprendre les méthodes d'imputation et de les appliquer avec précaution pour assurer l'intégrité des données et l'exactitude des résultats d'analyse