Techniques avancées d'imputation

Mamadou Mbodj, Septembre 2024
Contact : mamadou.mbodj@ansd.sn

Cours : Traitement de données

(ENSAE 2024)

Classe : ISE 3

ensae

ANSD
Agence Nationale de la Statistique et de la Démographie

1. Objectifs du cours

- Comprendre les différents types de données manquantes et savoir les identifier
- Parvenir à imputer avec des techniques avancées

2. Définition des types de données manquantes

- **Missing completely at Random**
  - Pas de biais statistique. Les données sont manquantes de façon totalement aléatoire, indépendamment de toutes les autres variables. Peut-être corrigé avec des méthodes d'imputation simples.
  - Exemple : Lors d'une enquête, certaines personnes oublient de remplir un champ, mais cela n'a rien à voir avec leur profil.

- **Missing at Random**
  - Les valeurs manquantes dépendent d'autres variables observées, mais pas de la variable elle-même. Peut être corrigé par l'imputation basée sur des modèles si les variables explicatives sont bien modélisées.
  - Exemple : Dans une enquête sur les revenus, certaines catégories sont plus susceptibles de ne pas répondre à la question sur le salaire.

2. Définition des types de données manquantes

- **Missing not at Random**
    - La probabilité de non-réponse dépend de la valeur elle-même (biais de non-réponse). Problématique majeure, car elle peut introduire un biais important dans les analyses.
    - Exemple : Les personnes ayant des revenus très élevés ou très faibles peuvent être plus enclines à ne pas déclarer leur salaire.

3. Comment identifier le type de données manquantes ?

- **Méthode d'analyse descriptive (MCAR vs autres)**

  - Taux de non réponses par variable
  - Vérification des patterns de valeurs manquantes

  Exemple : Si les valeurs manquantes se répartissent uniformément dans l'échantillon, elles sont probablement MCAR. Si elles sont concentrées sur un sous-groupe, elles sont plutôt MAR ou MNAR.

3. Comment identifier le type de données manquantes ?

❖ **Test de Little (MCAR vs autres)**

Le test de Little est un test statistique utilisé pour déterminer si les données manquantes sont manquantes complètement au hasard (MCAR). Il est basé sur la comparaison des moyennes des variables avec et sans données manquantes. Il repose sur une statistique du Chi-deux calculée en regroupant les données en classes selon les valeurs manquantes.

- H_0 (Hypothèse nulle) : Les données manquantes suivent un mécanisme MCAR (Missing Completely At Random).
- Diviser les données en groupes selon les types d'individus avec données manquantes.
- Comparer les moyennes des variables entre ces groupes.
- Calculer une statistique de test basée sur un Chi-deux ajusté.
- Si la p-value est faible (< 0.05), les données ne sont pas MCAR.

3. Comment identifier le type de données manquantes ?

- **Comparaison de distributions (MAR vs MNAR)**

  - Comparer la distribution des autres variables entre les individus avec et sans valeurs manquantes
    - Si les distributions sont similaires → MCAR ou MAR
    - Si elles sont différentes → Potentiellement MNAR

  Exemple : Si les non-répondants à la question du salaire sont surtout ceux dans les catégories à faible revenu, la non-réponse est MNAR.

- **Modélisation de la non réponse (MAR vs MNAR)**

  - Construire un modèle de classification pour prédire la probabilité d'absence d'une valeur
    - Si les variables explicatives influencent la non-réponse → MAR.
    - Si la variable manquante est prédictive d'elle-même → MNAR

4. Imputation par modélisation

- **Approche générale**
  - Imputation par prédiction
  - Capable de capturer des relations complexes entre les variables.
  - Plus performant que les méthodes traditionnelles dans certains cas.

- **Méthodes courantes**
  - Plus proches voisins
  - Régressions linéaire, polynomiale, etc. pour les variables quantitatives
  - Méthodes de classification pour les variables catégorielles (RF, Logistique, Arbres de décision, etc.)

5. Multiple Imputation by Chained Equations - MICE

- **Principe de l'imputation multiple :** Remplacer les valeurs manquantes par plusieurs imputations (plutôt qu'une seule), pour mieux refléter l'incertitude liée aux données manquantes.

- **Idée clé :**
  - Générer plusieurs jeux de données complets en remplaçant les valeurs manquantes par différentes imputations.
  - Effectuer l'analyse statistique sur chaque jeu de données.
  - Agréger les résultats pour obtenir des estimations robustes.

- **Pourquoi ne pas faire une seule imputation ?**
  - Une seule valeur imputée sous-estime la variabilité des données.
  - L'imputation multiple permet de tenir compte de l'incertitude liée à la non-réponse.

5. Multiple Imputation by Chained Equations - MICE

❖ **Principe de l'imputation multiple**

- Choisir un modèle d'imputation (ex. régression linéaire pour des variables continues, logistique pour des catégoriques).
- Imputer chaque variable manquante conditionnellement aux autres variables. Schéma du processus :
  - ✓ Imputation de Var1 en fonction des autres variables (Var2, Var3, …).
  - ✓ Imputation de Var2 en fonction de (Var1, Var3, …).
  - ✓ Etc.
- Répéter le processus plusieurs fois (itérations) pour converger vers une meilleure estimation.

6. Cas pratique