Les imputations du fichier de l'enquête Famille et logements (EFL 2011)

Programme Des lieux aux liens (LiLi),

Groupe d'exploitation de l'enquête,

Ined, 22 avril 2013

Laurent Toulemon

Plan de la présentation

- Pourquoi imputer les non réponses ?
- Les règles d'imputation
- Quelques exemples
- Le fichier « imputé »

I - Pourquoi imputer les non réponses ?

- 1) Le problème
- 2) Est-on intéressés par les non réponses ?
- 3) L'apurement du fichier

I - Pourquoi imputer les non réponses ?

- 1) Le problème
  - →20% de bulletins manquants
  - →Les bulletins de l'enquête
    - 0,7% de bulletins « presque vides » → Exclus
    - 72% de bulletins « presque pleins »
    - 18% de bulletins avec un bloc manquant
    - 5% avec deux, 2% avec 3, 2 avec plus de 3
  - →Les non-réponses sont-elles informatives ?
    - Sur la variable d'intérêt (refus de répondre, NSP)
    - Sur les variables explicatives (ou les contrôles)

2) Est-on intéressés par les non réponses ?

- Les non réponses sont informatives
  - →Ne pas répondre à la question sur la vie en couple, au fait d'avoir des enfants
- Les non réponses n'ont pas d'intérêt
  - →Oubli, problème de saisie
  - →Comportement, pas la manière de répondre
- L'apurement du fichier
  - →Simplifier l'analyse sans éliminer les « manquants »
  - →On « corrige » les erreurs triviales

3) L'apurement du fichier

- Vérification et validation
  - →Réalisées avec l'Insee
  - →Nécessaire car bulletins auto-remplis
- Réponses mal codées, mal saisies
  - →Réponses incohérentes (enfants nés avant Ego)
  - →Cas rares (couples de même sexe)
  - →Retour au bulletin
- Cas limites
  - →Pas d'enfant mais petits-enfants

II - Les règles d'imputation

- 1) Pourquoi une donnée est-elle manquante ?
- 2) L'imputation par hot deck
- 3) Les contraintes de l'imputation

II - Les règles d'imputation

- 1) Pourquoi une donnée est-elle manquante ?
  - →Réponse implicite : « pas concerné », « non », « zéro », → Imputation déterministe (apurement)
    - Frères et sœurs, demi-frères et demi-sœurs
  - →Réponse non informative → on « devine » la réponse à partir d'autres information du bulletin
    - Hot deck (donneur dans le fichier)
    - Cold deck (liste exogène, règle spécifique
    - Autres méthodes : estimation par régression
  - →Conserver les moyennes et les variances

2) L'imputation par hot deck

- Du « donneur » au « receveur »
  - →On cherche un « donneur », jumeau sur des variables pertinentes, qui a répondu à la question
    - dans un groupe de semblables
  - →Le bulletin « reçoit » la valeur du donneur
- Absence de biais, variance conservée
  - →Conditionnellement au groupe (pas dans l'ensemble)
- Cohérence des imputations
  - →Ne pas créer des incohérences nouvelles
  - →Prendre de nombreuses valeurs du même donneur

3) Les contraintes de l'imputation

- Trouver des donneurs en nombre suffisant dans chaque groupe
- Eviter que le même donneur ne serve pour plusieurs receveurs
- Utiliser autant que possible le même donneur pour chaque receveur
- Trouver des donneurs « complets » pour imputer plusieurs variables en même temps
- Conserver les associations qui nous intéressent

III – Quelques exemples

- Les frères et sœurs
- Mariage et Pacs
- Les enfants
  - →Présence d'enfants
  - →Les dates de naissance des enfants
  - →Les sexes des enfants
  - →La survie de enfants partis
- Les lieux de résidence des parents et des enfants

III – Quelques exemples (1/3)

- Les frères et sœurs
  - →Pas de raison de ne pas répondre
  - →Répartition correcte si NR implique « zéro »
- Mariage et Pacs
  - →Quand on est marié on n'est plus pacsé
  - →Comparaison avec le nombre de pacs : « manquent » des pacsés puis mariés
- Les enfants
  - →La partie sur les parents est-elle renseignée ?
  - →Et sur les petits-enfants ?

III – Quelques exemples (2/3)

- Les dates de naissance des enfants
  - →Trois dates « ancres » (18 ans, dernier couple, 2011)
  - →Nombre d'enfants d'Ego et du conjoint actuel
  - →Homothétie sur les dates pour « boucher les trous »
- Les sexes des enfants
  - →Le sexe des deux premiers dépend du nombre d'enfants du couple…
- La survie de enfants partis
  - →Non réponse autorisée en cas de décès
  - →Survie (tables de mortalité), puis lieu de résidence

III – Quelques exemples (3/3)

- Les lieux de résidence des parents et des enfants
  - →Proximité entre lieux (différentes variables)
  - →On impute deux variables
    - Proximité avec le lieu de résidence de Ego
    - Qualification du lieu (urbain – rural pour la France)
  - →A partir de ces variables et d'autres variables d'Ego, (Lieu de naissance de Ego) on impute le cadre (région en France, continent à l'étranger)
  - →Puis on impute un lieu par cold deck
    - Liste de pays ou de communes
  - →Et on peut ainsi créer toutes les variables de lieux

Conclusion

- Pas d'imputation « idéale »
- Le fichier « imputé »
  - →Variables imputées en plus
- Le programme d'imputation
  - →Réplicabilité, possibilité de modification (sauf pour les lieux)
- Une documentation
  - →Aspects théoriques et pratiques
- Merci pour vos questions