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, 12 juin 2013

Les étapes d'imputation

- Étude de la NR: par bulletin, par variable et par groupes de variables
- La désimputation du FPR
- Choix de l'ordre d'imputation
- Choix de la méthode d'imputation selon les variables
- Les critères de tri: quelles variables utiliser?
- Raffinement de la méthode par hot deck: éviter les effets de bord

- Les variables imputées: exemples

Étude de la NR: par bulletin, par variable et par groupes de variables

→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

→On regarde la proportion de données manquantes

- par variable (on utilise les variables indicatrices « I_ »)
- par groupes de variables: pour savoir si on impute par « paquets »

La désimputation du FPR

- Pourquoi?

Imputations non documentées

Qui créent parfois des incohérences

Imputations déterministes qui ne nous convient pas dans certains cas

Imputations partielles (on veut garder une même logique pour l'ensemble)

On supprime donc toutes les imputations SAUF pour la question q5sexe_c (2663 cas où i_q5sexe_c=2: liste ABC et données EAR)

Exemples:

- Incohérences:

Q1anai. L'Insee a pris l'année du recensement brute (1) ou imputée (2) . On refait une imputation si I_Q1anai = 2.

- Imputations déterministes

Q7pacs : Non si NR

Sous enregistrements par rapport aux données du Ministère.

- Q9tps_logr=1 si NR (Tout le temps si cohabitant)

Exemples (suite)

- Q11enfan. imputation selon petits enfants
- Q19 et Q25: imputations non documentées
- Q27grp. imputation non documentée.

- Dans le fichier enfant on supprime les enfants imputés (imputation non documentée 2614 enfants)

Choix de l'ordre d'imputation

Pourquoi?

On a besoin de certaines variables pour en imputer d'autres.

Globalement, c'est l'ordre du questionnaire qui est respecté.

Regroupement éventuel de certaines variables:

Exemple des questions 6, 7 et 8

- On doit garder les cohérences entre la mise en couple, le pacs et le mariage

Cohérence d'événements

Cohérence de dates

Donc on impute en créant une variable concaténée.

Choix de la méthode d'imputation selon les variables

- Réponse implicite : « pas concerné », « non », « zéro », → Imputation déterministe

Frères et sœurs, demi-frères et demi-sœurs (question 2)

- Réponse non informative → on « devine » la réponse à partir d'autres information du bulletin
  - Hot deck (donneur dans le fichier)
  - Cold deck (l'information vient d'ailleurs)

Imputation par hot deck

- Idée: on utilise l'information disponible pour inventer une information pertinente et probable.
- Méthode principalement utilisée dans notre programme.

Permet :

- L'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
  - → Éviter de « normaliser » les réponses (comme pour une imputation déterministe ex q9)

Cold deck

- On va chercher l'information ailleurs, soit dans un autre fichier, soit à partir de régularités connues.

Exemple: utilisation des tables de mortalité pour la survie des enfants du tableau 14

Hot deck: les critères de tri: quelles variables utiliser?

- Au départ, on tri par l'identifiant CABEFLA
  - →Réplicabilité, possibilité de modification (sauf pour les lieux)
- Pour faire des groupes le plus homogènes possible, on choisi les variables de tris

Quelle(s) variable(s) utiliser systématiquement ?

Sexe/Age/Cs

Création des variables groupesimput, groupesimput2 et groupesimput3

- On rajoute éventuellement une ou deux variables car liaison fonctionnelle ou statistique

Attention à avoir assez de donneurs…

Concrètement sur sas…

- On classe les individus (proc sort)
- Quand l'information que l'on cherche à imputer existe pour un individu, on met la valeur en réserve (donneur) (retain)
- Quand la valeur est manquante pour un individu on pioche dans la réserve. (receveur)

Attention à l'ordre de tri…

- Proc sort; by

Sexe age / age sexe

sexe	age	couple
1	22	1
1	22	2
1	23	?
1	23	2
1	23	1
2	22	1
2	22	2
2	22	2
2	23	1
2	23	2
2	23	1

age	sexe	couple
22	1	1
22	1	1
22	2	?
22	2	2
22	2	2
23	1	1
23	1	1
23	1	2
23	2	2
23	2	2
23	2	1

- Avec le premier tri (by sexe age) on a moins de cas où deux variables sont différentes, mais c'est un grand saut : la première femme (très jeune) sera éventuellement imputée avec la valeur du dernier homme (très vieux) ; avec le deuxième tri on a beaucoup de cas de petits sauts : le premier homme de chaque âge est imputé avec une femme d'un an de moins.

Raffinement de la méthode: éviter les effets de bord

- Éviter autant que possible que le premier d'un groupe soit manquant et que deux manquant se suivent. Éviter que le même donneur ne serve pour plusieurs receveurs

- Création de la variable NUM

Exemple, programme sas

- Questions 7 et 8:

```
/* On impute pacs ET mariage lorsque les deux sont en nr.*/
Proc sort data=x.testimput ; By q3couple_i q5sexe_c groupesimput3; Run;
data x.testimput;
set x.testimput;
By q3couple_i q5sexe_c groupesimput3;
retain n_ok n_miss pb;
if first.groupesimput3 then do;
   n_ok=1; n_miss=2;
end;
if q8mari not in ("Z"," ") and q7pacs not in ("Z"," ") then do; num=n_ok;
   n_ok=n_ok+2; end;
else if q8mari in ("Z"," ") and q7pacs not in ("Z"," ") then do; num=n_miss;
   n_miss=n_miss+2; end;
else kk=1;
if last.groupesimput3 then do;
   if n_ok < n_miss then pb = 1;
end;
run;
```

```
Proc sort data=x.testimput ;
By q3couple_i q5sexe_c groupesimput3 num;/*on ajoute num pour eviter les effets de
    bord*/
Run;
data x.testimput;
set x.testimput;
retain reserve;
retain reserve2;/*retain est fait sur le vecteur: la reserve*/
/*si il y a une reponse à mari on remplit la réserve*/
if q8mari not in ("Z"," ") and q7pacs not in ("Z"," ") then do;
reserve=q8mari;
reserve2=q7pacs;
end;
/*si les réponses à année mari et année pacs sont vides on va piocher dans la réserve*/
if pacsmari="ZZ" then do;
q8mari_i=reserve;
q7pacs_i=reserve2; /*on impute*/
end;
drop reserve reserve2 num n_miss n_ok pb kk;
run;
```

Les dates de naissance des enfants

- Les dates de naissance des enfants

On fixe des rangs

Les enfants des deux conjoints sont nés après les enfants d'un seul conjoint

Les enfants partis sont plus vieux que les enfants du logement

On crée un fichier avec les enfants en colonnes

Conserver les cohérence des dates…

→« ancres » temporelles (18 ans de ego et du cjt, année d'enquête, dernier couple (année de mise en couple ou de rupture)

ancres si en couple actuellement :
enfant de Ego avant couple : majoriteego - datecouple
enfant de cjt avant couple : majoritecjt - datecouple
enfant de couple : datecouple - anenquete
ancres si pas en couple actuellement :
enfant de Ego avant couple : majoriteego – anneerupture

→Type d'événement précédent rempli, puis suivant

→On utilise des donneurs complets

→Nombre d'enfants du couple, d'Ego, du conjoint actuel

→Homothétie sur les dates pour « boucher les trous »

Les variables imputées

- Variables en plus: avec le suffixe _i
- Ensemble des variables de la page 1, q11 et 12

Et années de naissance des enfants

- Reste à faire…

- Aucune imputation « parfaite »

Merci pour vos questions