Sélection, redressement, imputations: les dernières tendances en matière de multimode

Stéphane Legleye, DRTI

Plan

- I. Le cadre général

  Notation, effet de sélection et de mesure

- II. Mesurer l'effet – mode ou corriger le biais de sélection

  Les différentes méthodes

- III. Que faire lorsque l'effet de mesure est nuisible?

  Une méthode pragmatique pour contenir l'effet de mode en corrigeant l'effet de mesure

- IV. Un exemple

  Les « opinions » dans CAMME (10 variables catégorielles)

I. Les enquêtes multimode (1)

- Deux modes de collecte/enquêtes A et B produisant 2 échantillons de répondants A et B

- Par convention A est le mode de référence
  - historique ou réputé plus adapté : thématique, population
- Populations A et B recouvrantes
  - Exclusion des cas disjonctifs (type A pour les jeunes, B pour les vieux)

I. Les enquêtes multimode (2)

1. A et B attribués en randomisé

2. Protocole séquentiel : A imposé puis B si non-réponse

3. Protocole au choix simple: proposition de A ou B

4. Protocole au choix séquentiel : proposition de A ou B, relance sur le mode alternatif si non-réponse

5. Deux enquêtes probabilistes différentes, A et B

6. Deux enquêtes, A probabiliste, B non probabiliste

I. Définition des variables / questionnaires

- X les variables sociodémographiques ou de la base de sondage
- Y les variables d'intérêt

I. Sélection et mesure

- Effet de mode=biais de sélection + biais de mesure
- Biais de sélection
  - Deux ensembles de répondants

    => Distributions de X diffèrent dans A et B
  - Vrai même en cas d'expérimentation randomisée (sauf en laboratoire ou si taux de participation ≈ 100 %)
- Biais de mesure (résiduel) ou effet de mesure
  - Deux modes/enquêtes

    => Distributions de Y *conditionnellement à X* diffèrent dans A et B

I. L'effet de mesure

du point de vue conceptuel

Un individu répond à un unique questionnaire sur un unique mode.

- $i \in A \cup B$, $Y_i(1)-Y_i(0)$, valeur dans B – valeur dans A
- seule une valeur est observée, l'autre est **contrefactuelle**

- Souvent mesuré par l'« effet moyen du traitement » (ATE) :
  - $E[Y_i(1)-Y_i(0)|X]$, $i \in A \cup B$

L'effet de mesure est un problème de réponse partielle causé par le mode

Donc un problème d'imputation

I. Les causes du biais de sélection

- Toute enquête est multimode : annonce, contact et interview-passation sont réalisés sur des modes différents
- Biais varie suivant :
  - Mode d'annonce et de contact (postal, téléphonique, courriel…)
  - Délai entre annonce, contact et questionnement
  - Caractère officiel, crédible, institutionnel de l'annonce et du contact
  - Thème de l'enquête
  - Mode de vie - disponibilité du ménage / personne sélectionnée
  - Compétences des personnes sélectionnées, préférence pour un mode
  - Décision de participer est un arbitrage coût / intérêt (Leverage-saliance theory)

I. Les causes du biais de mesure

.Présence d'un enquêteur : communication riche (verbale, paralinguistique, visuelle - cartes ou bien ergonomie du questionnaire- etc.) et stimulante

- encourage le répondant, reformule les questions, décourage les abandons, s'assure de la consultation des documents nécessaires etc.
- Peut brider les déclarations perçues comme stigmatisantes ou non-valorisantes : *biais de désirabilité sociale*

.Absence d'enquêteur : tout repose sur le support (verbal, visuel)

- encourage la réponse sincère sur des questions sensibles
- Mais peut diminuer la motivation, inciter à se contenter de réponses approximatives (*satisficing* - « satisfaisance »)

II. Estimer un effet de mode

II. Méthodes classiques

Biais de sélection

- Objectivation de l'équilibrage entre A et B pour les variables X
  - Khi²s, t-tests, statistiques d'associations
    - sensibles à la taille de l'échantillon
  - différences standardisées D et ratios de variance R:
    - Continues : $D = 100\ (m_1-m_0)/\sqrt{((s_1^2+s_0^2)/2)}$: ($|D|<10$)
    - Binaires : $D = 100\ (p_1-p_0)/\sqrt{((p_1q_1+p_0q_0)/2)}$ extensible aux catégorielles
    - Ratios de variances : $R=s_1^2 / s_0^2$: R proche de 1 ($0,5<R<2.0$)

II. Méthodes classiques

estimation de l'effet de mesure (1)

Repondération: rendre les distributions en X comparables en A et B pour réduire le biais de sélection

1. Calage de B sur A (ou sur la population cible) pour les X

2. Pondération par l'inverse du score de propension (IPW) :

   score= probabilité d'être dans B plutôt que A conditionnellement à X: Pr(B | X)
   Estimé par régression logistique

   - Poids : p_ipw=poids/score
   - Utilisation éventuelle de Groupes de réponse homogènes (GRH)

3. Combinaison IPW - calage

⇒ Estimation de l'effet de mesure par simple différence de moyenne de Y sur la repondération de {A, B} (ou modèles multivariés)

II. Méthodes classiques

estimation de l'effet de mesure (2)

Avantage de la repondération :

- on ne perd aucune observation (ou presque), donc généralisabilité maximale
- IPW (variables internes à l'enquête) et calage (source externe) combinables

Limite :

- variance accrue des poids
- conservation d'individus atypiques
- variance des estimateurs sous-estimée car la repondération est issue d'un modèle
- Biais potentiel si :
  - D et R restent trop grands pour le score ou certaines variables
  - R des résidus des variables x'j conditionnellement au score restent trop grands (<0.5 ou >2)

II. Méthodes classiques

estimation de l'effet de mesure (3)

- Appariement (matching): sur score de propension
  - Le score classe les observations de B et de A sur un seul critère
  - Equilibrage asymptotique des distributions de X dans les échantillons appariés

$\Rightarrow$ Estimation de l'effet de mesure par simple différence de moyenne de Y sur la repondération de {A, B} (ou modèles multivariés ajustant sur le score)

Avantages : comparaison d'individus comparables

- mesure d'un effet plus propre
- pas d'autre poids que le poids initial (pas de repondération)

Limite : perte de généralisabilité car appariement incomplet

- perte de puissance

II. Méthodes classiques

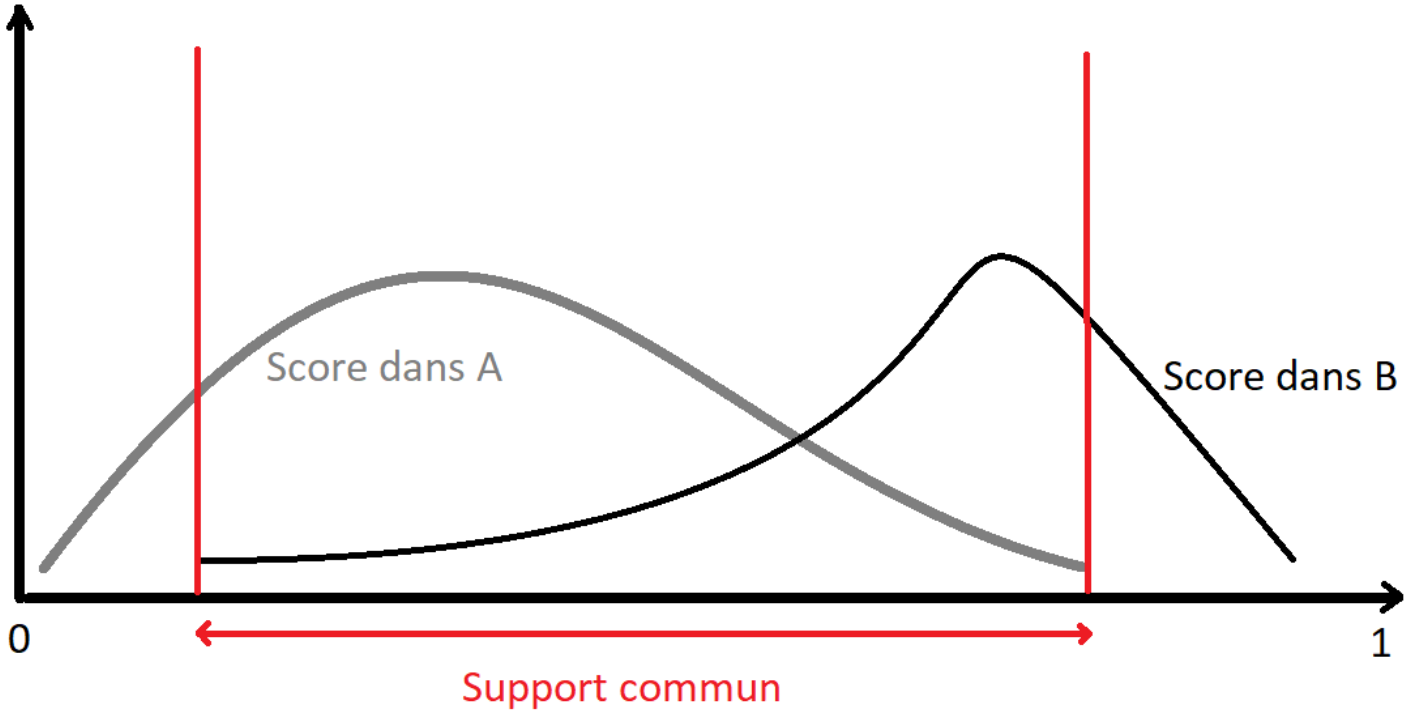
estimation conjointe du biais de sélection et de l'effet de mesure

- Dernière famille: méthodes de décomposition
  - Exemple type= Oaxaca-Blinder

- Hors sujet ici, car ne propose pas de correction

II. Le rôle central du score de propension

- Peut servir à déterminer le support commun
  - Caractéristiques X communes à des observations des deux modes A et B
  - Plage de scores commune aux deux échantillons
  - Décrire simplement les défauts de couverture de A et B en X



II. Le rôle central du score de propension

- La conservation d'individus atypiques (hors support commun ou très rares pour les calages) explique la hausse de la variance des poids des repondérations
- Quelles variables y inclure ?
  - Plus on met de variables de X, plus on équilibre les échantillons A et B en X : il faut calculer les D et R
  - mais on augmente la variance des poids

➔ Il y a des règles : (voir diapositives de fin)

- Défaut: la construction est itérative…

III. Que faire lorsque l’effet de mesure est nuisible?

III. La méthode proposée
Rappel

L'effet de mesure…

1. Peut être néfaste ou bénéfique pour la qualité de la mesure

   Cas de l'absence d'enquêteur, absence de désirabilité sociale…

2. Peut être jugé néfaste pour la diffusion des données

   Rupture de série, etc.

Ne pas le corriger peut être un choix

III. La méthode proposée
Métaphore

Situation: le diagnostic est posé (il y a un effet de mesure). Il est nuisible; c'est un effet de réponse. Comment traiter?

Chirurgie: ablation du Y puis greffe d'un Y plus conforme (imputation)

Traitement curatif de la dernière chance.

Idée naturelle: Imputer TOUT B par rapport à A (référence)…

Chirurgie militaire ancien régime…

Alternative: imputer les individus de B porteurs de l'effet de mesure.

Chirurgie plus moderne…

III. La méthode proposée
Principe

Deux types d'observations sont potentiellement concernées par un « effet de mode » :

1/ Celles qui diffèrent du point de vue des X (hors support commun): c'est le complément de couverture de B relativement à A

Elles portent le biais de sélection; les traiter peut être non-souhaitable.

2/ Celles qui sont dans le support commun mais dont les valeurs de Y diffèrent.

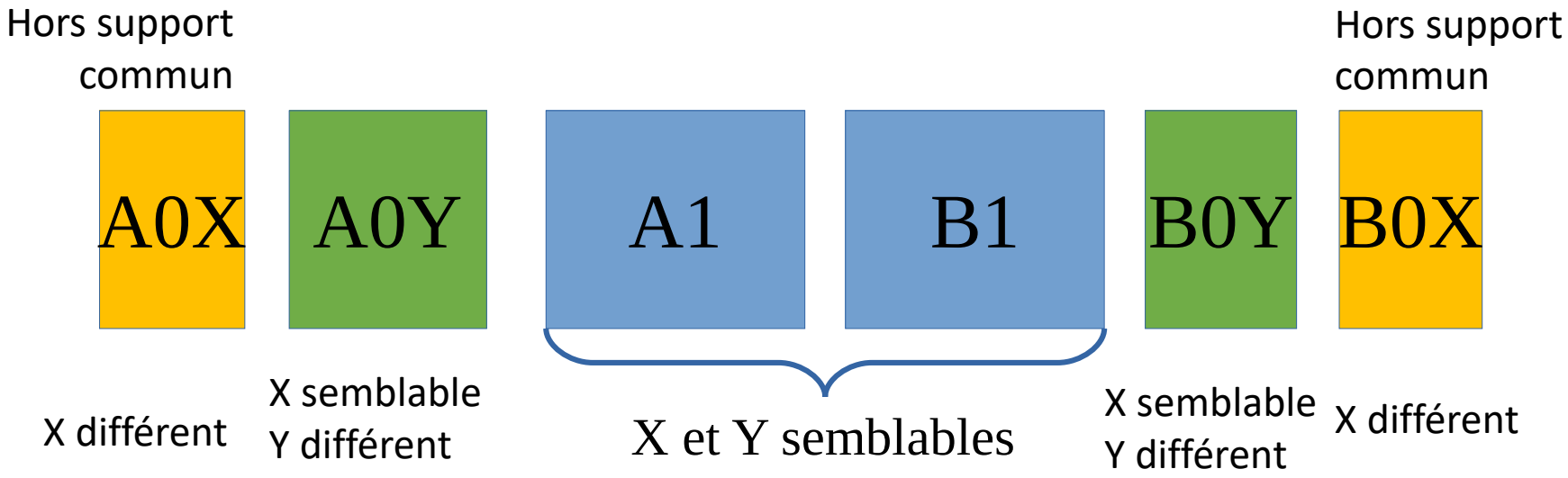
Elles portent un authentique effet de mesure : conditionnellement à X, leur Y diffère des Y mesurés dans B

On ne va traiter que ces dernières : correction de l'effet de mesure uniquement

III. Principe de la méthode proposée
Illustration

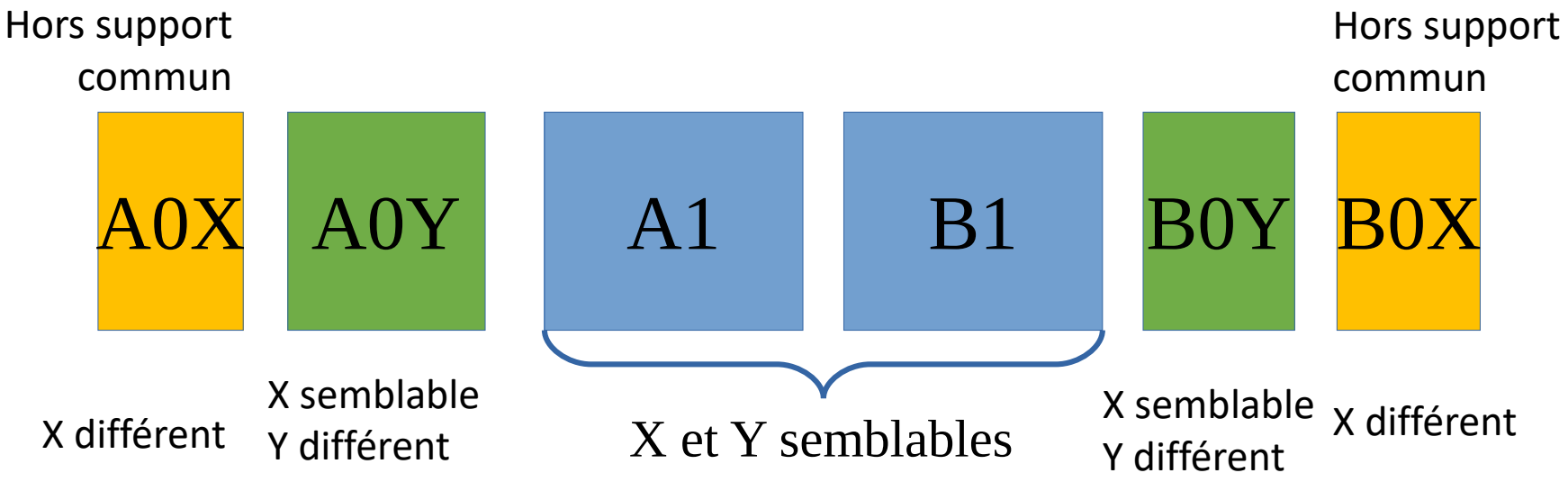
1/ Identification des répondants de A et B :

- similaires en X et Y : A1 et B1: jumeaux complets
- différant en X : A0X et B0X: porteurs de l'effet de sélection
- similaires en X mais différant en Y : A0Y et B0Y; porteurs de l'effet de mesure



III. La méthode proposée

2/ Imputation des Y de B0Y



III. Choix techniques

I. Appariement sur score de propension

1. Définition du support commun (➔ A0X et B0X)

   Repose sur un premier score de propension en X

2. Appariement sur le support commun en X et Y (➔ A0Y et B0Y)

   Repose sur un second score de propension en X et Y

   Le Y peut être utilisé brut ou transformé (quantiles pour les continues)

II. Imputation de B0Y

1. Suivant {A1, B1} ou {A0X, A0Y, A1, B1}…

IV. Exemple

Les « opinions » dans CAMME

Présentation

I. Mode A: Enquête téléphonique été 2015

Base de sondage : TH, restreinte aux ménages en liste blanche

N=4698, n=2782 répondants

II. Mode B : Enquête Internet en parallèle

Base de sondage : TH

N=105000, n=13119 répondants

Présentations des variables

Variables d’ « opinion » : n=10

Avis sur la conjoncture passée et future, intention d’achat de quelques biens d’équipement

Echelle de type Likert à 5 positions

Variables finales en 3 modalités ordonnées : -1, 0, 1

Ces variables ne forment **pas** un ensemble **unidimensionnel**:

Une ACP suggère que 2-3 dimensions sont utiles (valeurs propres 3.3, 1.3, 1.0, 0.95, 0.7)

CNRT téléphone

1. Age de la personne de référence (ageprcl, 5 classes)
2. Revenu du foyer en déciles (r_foy_rev)
3. Statut matrimonial (5 classes : mcdvt1)
4. Statut d'occupation du logement (4 classes, occ : propriétaire, locataire, gratuit, autre)
5. Type de logement (2 classes, natloc_r : maison, autre)
6. Région (21 classes, rg)
7. Taille d'unité urbaine (9 classes, tu)
8. téléphone fixe, téléphone mobile et courriel (8 classes, telfixemobemail)
9. Nombre d'appels (nbappel)
10. Nombre d'appels au carré (nbappelsq)

Modélisation logistique (aire sous la courbe ROC : AUC=0.76)
GRH : 11 classes

CNRT Internet

1. Age de la personne de référence du foyer (5 classes)
2. Revenu du foyer en déciles
3. Statut matrimonial (5 classes)
4. Statut d'occupation du logement (4 classes : propriétaire, locataire, gratuit, autre)
5. Type de logement (3 classes)
6. Région
7. Taille d'unité urbaine
8. Combinaison de la présence d'un téléphone fixe, d'un téléphone mobile et d'un courriel

Modélisation logistique (AUC=0.72)
GRH en 12 classes

Calage de l'internet sur le téléphone

On cale ensuite l'échantillon B sur A

3 variables pour les marges : sexe, âge et diplôme

Statistiques de poids :

- Téléphone : max/min=8.11, CV=46.6
- Internet : max/min=108.46, CV=104.7

Equilibrage

D_X=7.1 (mauvais pour nature du logement, statut d'occupation, quintiles de revenus)

D_Y=19.7, les dix variables ont un D>10

(Confirmé par des analyses multivariées)

Estimation de l'effet de mesure

2. Amélioration potentielle : CNRT puis repondération IPW de l'Internet sur le téléphone puis calage

Statistiques de poids pour Internet (téléphone inchangé):

max/min=719, CV=162.4 : augmentation de 60%

D_X=7.5

D_Y=18.3, huit variables ont un D>10

(Confirmé par des analyses multivariées.)

Le patient ne s'y opposant pas, on tente la chirurgie

On reste sur la pondération CNRT + calage

1/Détermination du support commun

Premier score de propension

Toutes les variables sociodémographiques sont prises en compte : n=10

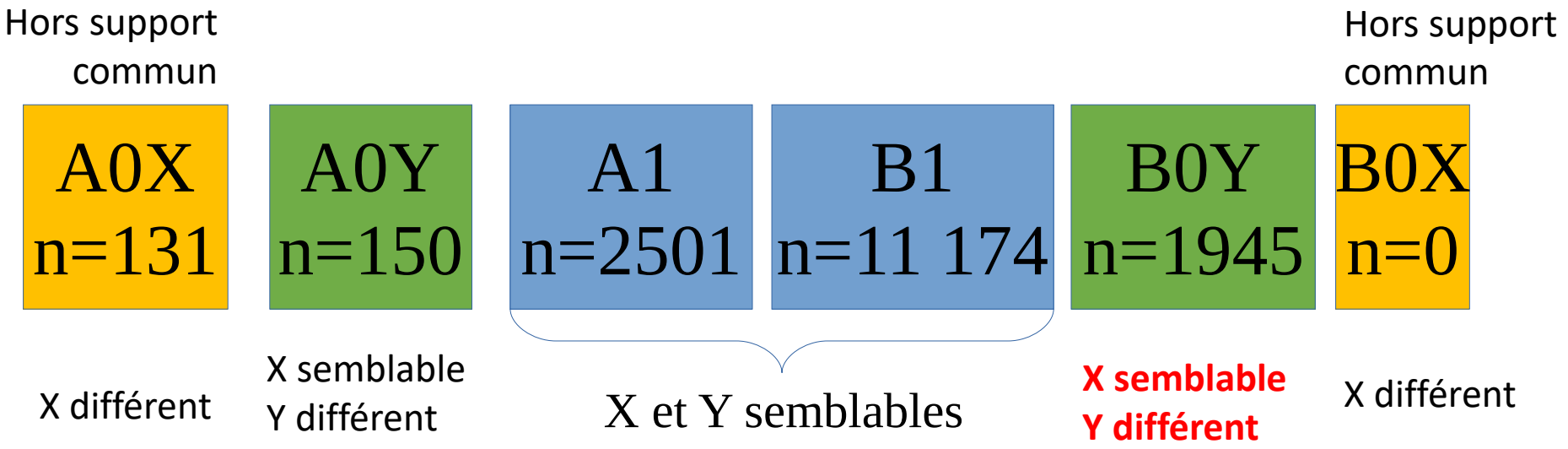
sexe_r, ageprcl_rr, diplome3c_r, occupa, mcdvt1_r, natloc_r, occ, r_foy_rev_rr, nbpers_rr, idf, tuu)

131 observations hors support commun sur téléphone (A0X), aucune sur Internet (B0X est vide)

2/ Appariement sur X et Y

Sur le support commun:

second score de propension incluant 21 variables (les 11 X et les 10 Y)



B0Y représente 14.8% de B

2/ Appariement sur X et Y

Equilibrage {A1, B1}:

D_X=3.1 : très bon

D_Y=11.1 (6 D>10) : très amélioré

Equilibrage {B0Y, A1-B1}:

D_X=12.2 : dégradé

D_Y=50.1 (10 D>10) : très mauvais

3/ Imputation des Y de B0Y

On choisit d'imputer à partir des observations {A, B1}

(on n'élimine pas les A0X ni les A0Y)

Construction du modèle avec sélection automatique des variables

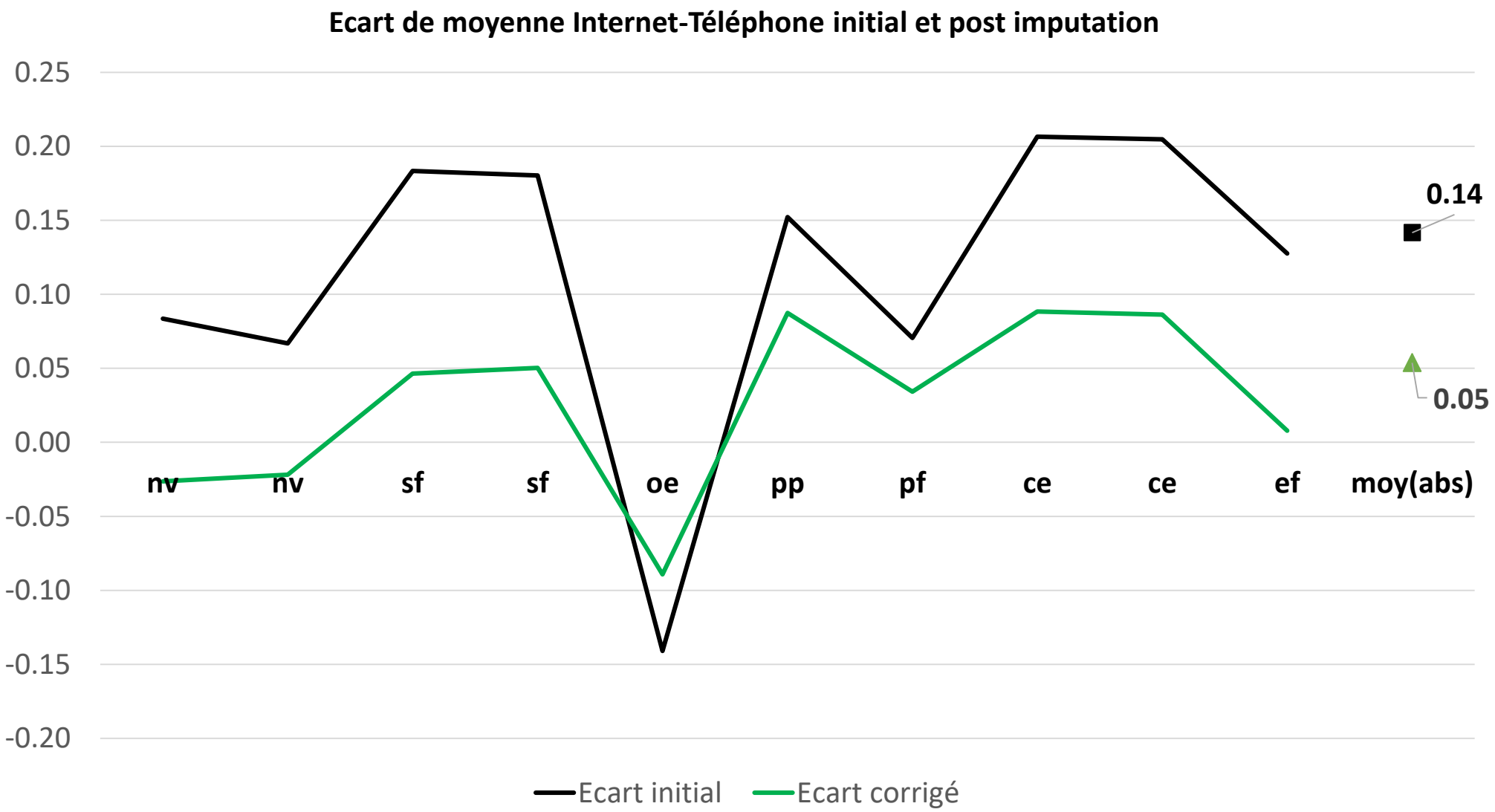
- (modélisation linéaire multivariée et validation croisée) sur les 3 premières composantes principales de l'ACP de Y

Imputation type hotdeck aléatoire de chacune des 10 variables de Y

Equilibrage post imputation :

- D_X inchangé=7.1
- D_Y=7.0 (2D>10) **vs 19.7 (10D>10)** : réduction de 64%

Variable par variable en différence de moyenne



Intérêts (1)

- Cible et corrige l’effet de mesure uniquement
- Fonctionne quand la réduction du biais de sélection est insuffisante
- N’accroît pas la variance des poids
- Méthodes économétriques simples et automatisables
- Qualité de l'appariement en X et Y dépend a priori « peu » du modèle si l'équilibrage est bon
- Possibilité de prendre des composantes ACP des Y : compatible avec enquête très multithématiques

Intérêts (2)

- Fonctionne pour tout type de protocole et de configuration A et B
  - Dès lors qu'une pondération prend en charge la « représentativité »
  - Par exemple: si le papier est proposé après refus Internet, on aurait une pondération reflétant cette sélection après une première phase de CNRT ; le reste serait à l'identique

Limites

- On impute les données… et cela peut concerner beaucoup d'individus
- Si Y large, correction des effets sans doute plus réduite
- si Y est multiblocs (multithématique avec filtrage), complexité accrue, car il faut stratifier les opérations par filtre
- Il faut bien choisir les variables du modèle de propension: problème des facteurs de confusion inobservés
- Efficacité potentiellement plus limitée si Y binaire
- Imputation ➔ variance (calculable en imputations multiples)
- Diffusion : imputations multiples ou bien choix d'un jeu d'imputation
- Utilité doit se tester au cas par cas

Annexes

Règles de constitution du score

- **Inclure tous** les facteurs de confusion (variables associées au mode de collecte et à Y) primaires y compris interactions : pas de médiateurs (facteur de confusion affecté par le mode)
- Hors facteurs de confusion : pas de biais mais inclure
  - variables associées à Y mais pas au mode décroît la variance de l'estimateur d'effet
  - variables associées au mode mais pas à Y accroît la variance de l'estimateur d'effet
- Ne pas utiliser de sélection automatique : pourrait manquer des facteurs de confusion. Donc : littérature et analyse préalable, diagrammes causaux (Rubin 2007)

Les limites du score

- On perd les individus dont le score est 0 ou 1
- Impossible d'être sûr de la qualité du modèle
  - Variables non mesurées, vrai mécanisme inconnu
  - Les statistiques classiques d'adéquation ne sont qu'un indice (Lackfit, R^2, aire sous la courbe ROC…)
- Mais le meilleur modèle sera celui :
  - qui équilibre le mieux X dans les échantillons A et B
  - qui ne déforme pas trop les poids

=> donc construction itérative

Les alternatives au score

- Alternatives à la méthode itérative du score de propension :
  - Estimateur BEAST (balancing equation algorithm with selection for treatment effect) de Bléhaut et al. (document de travail du CREST) : non testé
    - Mesure de l'ATE sur échantillon calibré résolvant le choix de la sélection de variables...
  - Entropy balancing (Hainmueller 2012) : non testé
  - Coarsened exact matching (King et al. 2011) : en cours de test

Références

Razafindranovona T, La collecte multimode et le paradigme de l'erreur d'enquête totale, Document de travail M 2015/01, Insee

Rubin DR, The design vs the analysis of observational studies for causal effect: Parallels with the design of randomized trials, Statistics in medicine, 2007, 26:20-36

Austin PC, Stuart EA, Moving toward best practice whein using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies, Statistics in medicine, online, 2015

Verzillo S, Berta P, Bossi M, %CEM, a SAS macro to perform Coarsened exact matching, Journal of statistical computation and simulation, online, 2016

Bléhaut M, D'Hautefoeuille X, L'Hour J, Tsybakov A, A calibration estimator for treatment effect and synthetic control in high-dimension, document de travail Crest, en cours, 2016

Hainmueller J, Entropy Balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies, Political analysis, 2012 20:25-46