



**APPROCHE PAR MODÈLE DE  
NON-RÉPONSE POUR  
L'INFÉRENCE EN PRÉSENCE DE  
DONNÉES IMPUTÉES**

**DAVID HAZIZA & J.N.K. RAO**

**Statistique Canada & Université Carleton**

**Paris**

**15 Mars 2004**

# Plan de la présentation

- Introduction
- Imputation par la régression
- Estimateur imputé
- Approches pour l'inférence
- Imputation par la régression modifiée
- Conclusions

# Contexte

- Population finie de taille  $N$
- L'objectif est d'estimer le total dans la population

$$Y = \sum_{i \in U} y_i,$$

pour une variable d'intérêt  $y$ .

- On tire un échantillon aléatoire,  $s$ , de taille  $n$ , selon un plan de sondage  $p(\cdot)$ .

# Estimation: réponse complète

- Un estimateur de  $Y$  est l'estimateur de Horvitz-Thompson

$$\hat{Y} = \sum_{i \in s} w_i y_i,$$

- $w_i = 1/\pi_i$  désigne le poids de sondage de l'unité  $i$
- $\pi_i$  désigne la probabilité d'inclusion de l'unité  $i$  dans l'échantillon  $s$ ;  $i = 1, \dots, N$ .

# Non-réponse: estimateur imputé

- En présence de non-réponse à la variable  $y$ , on définit un **estimateur imputé** de  $Y$  selon

$$\hat{Y}_I = \sum_{i \in S} w_i a_i y_i + \sum_{i \in S} w_i (1 - a_i) y_i^*,$$

où  $a_i$  est une variable indicatrice de réponse telle que

$$a_i = \begin{cases} 1 & \text{si l'unité } i \text{ a répondu à la variable } y \\ 0 & \text{sinon} \end{cases}$$

et  $y_i^*$  désigne la valeur imputée utilisée pour remplacer la valeur manquante  $y_i$

# Imputation par la régression déterministe

- Soit  $\mathbf{z}_i = (z_{1i}, \dots, z_{qi})'$  un vecteur de **variables auxiliaires** disponibles pour toutes les unités échantillonnées
- Les valeurs imputées peuvent être obtenues en ajustant le modèle de régression

$$m: y_i = \mathbf{z}_i' \boldsymbol{\beta} + \varepsilon_i, \\ E_m(\varepsilon_i) = 0, E_m(\varepsilon_i \varepsilon_j) = 0, i \neq j, V_m(\varepsilon_i) = \sigma^2 \boldsymbol{\lambda}' \mathbf{z}_i \equiv \sigma^2 c_i^{-1}$$

au moyen des unités répondantes.

Valeurs imputées:  $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r,$

$$\hat{\mathbf{B}}_r = \left( \sum_{i \in S} w_i a_i c_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in S} w_i a_i c_i \mathbf{z}_i y_i$$

# Estimateur imputé

- L'estimateur imputé devient:

$$\hat{Y}_I = \hat{Y}_r + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \hat{\mathbf{B}}_r$$

$$\hat{Y}_r = \sum_{i \in S} w_i a_i y_i$$

$$\hat{\mathbf{Z}} = \sum_{i \in S} w_i \mathbf{z}_i$$

$$\hat{\mathbf{Z}}_r = \sum_{i \in S} w_i a_i \mathbf{z}_i$$

- Similaire à un estimateur GREG dans le cas de l'échantillonnage à deux phases

# Mécanisme de non-réponse

- Le comportement de réponse/non-réponse est vu comme un phénomène aléatoire
- La distribution des variables indicatrices de réponse,  $P(a_i | s)$ , est appelée **mécanisme de non-réponse**

$$a_i \stackrel{ind.}{\sim} B(1, p_i)$$

où  $p_i = P(a_i = 1 | s; i \in s)$  désigne la probabilité de réponse pour l'unité  $i$

# Mécanisme de non-réponse

On distingue 3 types de mécanisme de non-réponse:

1. **Mécanisme uniforme :**

$$p_i = p$$

2. **Mécanisme non-confondu :** la probabilité de réponse peut dépendre d'un vecteur de variables auxiliaires  $\mathbf{x}$  mais pas de la variable d'intérêt  $y$

$$p_i = P(a_i = 1 | y, \mathbf{x}) = P(a_i = 1 | \mathbf{x})$$

3. **Mécanisme confondu :**

$$p_i = P(a_i = 1 | y, \mathbf{x})$$

# Approches pour l'inférence

- Traditionnellement, deux approches ont été utilisées dans la littérature pour l'inférence en présence de données imputées:

1. Approche par Modèle de Non-Réponse Uniforme (AMNRU) (Rao, 1990)

Le mécanisme de non-réponse est uniforme,  $p_i = p$

2. Approche par Modèle d'Imputation (AMI) (Särndal, 1992)

Le mécanisme de non-réponse est non-confondu et on fait appel à un modèle d'imputation

$$m : y_i = \mathbf{z}_i' \boldsymbol{\beta} + \varepsilon_i, \\ E_m(\varepsilon_i) = 0, E_m(\varepsilon_i \varepsilon_j) = 0, i \neq j, V_m(\varepsilon_i) = \sigma^2 \boldsymbol{\lambda}' \mathbf{z}_i \equiv \sigma^2 c_i$$

# Une troisième approche

## Approche par Modèle de Non-Réponse Généralisé (AMNRG)

Le mécanisme de non-réponse est non-confondu et on fait appel à un modèle de non-réponse

$$\log \frac{p_i}{1 - p_i} = \mathbf{x}_i' \boldsymbol{\gamma}$$

où  $\mathbf{x}_i$  est un vecteur de variables auxiliaires disponibles pour toutes les unités échantillonnées.

**Remarque:** L'AMNRU est un cas particulier de l'AMNRG

# Quelle approche utiliser?

Modéliser la variables d'intérêt?

ou

modéliser la probabilité de réponse?

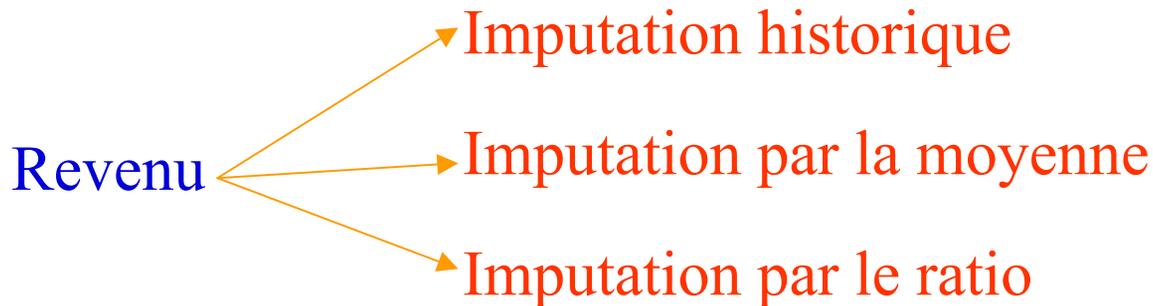
ou

modéliser les deux?

# Exemples

## (1) Projet de la TPS

Plusieurs méthodes d'imputation pour une même variable



## (2) Enquête sur les dépenses et immobilisations

Montant dépensé pour l'équipement par les compagnies :

Beaucoup de 0!

# Décomposition de l'erreur totale

- L'erreur totale,  $\hat{Y}_I - Y$ , peut être décomposée comme suit:

$$\underbrace{\hat{Y}_I - Y}_{\text{erreur totale}} = \underbrace{(\hat{Y} - Y)}_{\text{erreur due à l'échantillonnage}} + \underbrace{(\hat{Y}_I - \hat{Y})}_{\text{erreur due à la non-réponse}}$$

- Nous évaluons les propriétés, étant donné l'échantillon  $s$
- Cela permet de mettre l'accent sur l'erreur de non-réponse

➔  $\hat{Y} - Y$  est non aléatoire

# Biais de non-réponse sous l'AMNRU

$$\text{Biais}(\hat{Y}_I | s) = E_r(\hat{Y}_I - \hat{Y} | s) \approx 0$$

- Sous l'AMNRU,  $E_r(a_i | s) = p$

# Biais de non-réponse sous l'AMI

$$\text{Biais}(\hat{Y}_I | s) = E_r E_m(\hat{Y}_I - \hat{Y} | s) \approx 0$$

- Sous l'AMI,  $E_m(y_i | s) = \mathbf{z}_i' \boldsymbol{\beta}$

# Biais de non-réponse sous l'AMNRG

$$\text{Biais}(\hat{Y}_I|s) = E_r(\hat{Y}_I - \hat{Y} | s) = -\sum_{i \in s} w_i(1-p_i)(y_i - \mathbf{z}'_i \hat{\mathbf{B}}_p) \neq 0$$

• Sous l'AMNRG,  $E_r \left( \frac{\sum_{i \in s} w_i p_i c_i \mathbf{z}_i y_i}{\sum_{i \in s} w_i p_i c_i} \right)^{-1}$

$$\hat{\mathbf{B}}(\hat{Y}_I|s) = -\sum_{i \in s} w_i a_i \frac{(1-p_i)}{p_i} (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r)$$

# Comparaisons

	Approche par modèle de non-réponse uniforme	Approche par modèle d'imputation	Approche par modèle de non-réponse généralisé
$\hat{Y}_I$ avec $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$	Sans biais	Sans biais	Biaisé

# Estimateur ajusté

- On définit un estimateur ajusté par

$$\hat{Y}_I^a = \hat{Y}_I - \hat{B}(\hat{Y}_I | s)$$



$$\hat{Y}_I^a = \sum_{i \in S} w_i \mathbf{z}'_i \hat{\mathbf{B}}_r + \sum_{i \in S} w_i \frac{a_i}{p_i} (y_i - \mathbf{z}'_i \hat{\mathbf{B}}_r)$$

- Similaire à un estimateur par la régression
- En pratique, les  $p_i$  ne sont pas connues  $\longrightarrow$  utiliser  $\hat{p}_i$

# Caractéristiques

- Si  $\hat{p}_i \approx p_i$ , l'estimateur ajusté est approximativement sans biais sous l'**AMNRG**
- Si le modèle d'imputation est correctement spécifié, l'estimateur ajusté est approximativement sans biais sous l'**AMI**

 Valide sous les deux approches

**Désavantage:** requiert les variables indicatrices de réponse  $a_i$  ainsi que les probabilités de réponse  $\hat{p}_i$  dans le fichier

# Comparaisons

	Approche par modèle de non-réponse uniforme	Approche par modèle d'imputation	Approche par modèle de non-réponse généralisé
$\hat{Y}_I$ avec $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$	Sans biais	Sans biais	Biaisé
$\hat{Y}_I^a$ avec $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$	Sans biais	Sans biais	Sans biais

# Imputation par la régression modifiée

- Le but est de **déterminer des valeurs imputées** telles que l'estimateur imputé,  $\hat{Y}_I$ , soit sans biais pour  $Y$
- Nous cherchons des valeurs imputées de la forme,  $y_i^* = \mathbf{z}_i' \boldsymbol{\beta}$ , où  $\boldsymbol{\beta}$  est supposé connu pour l'instant

$$\longrightarrow \hat{Y}_I = \sum_{i \in S} w_i a_i y_i + \sum_{i \in S} w_i (1 - a_i) \mathbf{z}_i' \boldsymbol{\beta},$$

- On détermine  $\boldsymbol{\beta}$  tel que

$$\text{Biais}(\hat{Y}_I | s) = E_r(\hat{Y}_I - \hat{Y} | s) = 0$$

# Imputation par la régression modifiée

$$\text{Biais}(\hat{Y}_I | s) = 0 \Leftrightarrow \tilde{\mathbf{B}} = \left( \sum_{i \in s} w_i (1 - p_i) c_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in s} w_i (1 - p_i) c_i \mathbf{z}_i y_i$$

$\tilde{\mathbf{B}}$  ne peut être calculé puisque certaines valeurs de  $y$  dans  $s$  sont manquantes et que les probabilités de réponse  $p_i$  ne sont pas connues

$$\begin{aligned} \longrightarrow \tilde{\mathbf{B}}_r &= \left( \sum_{i \in s} w_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} a_i c_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in s} w_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} a_i c_i \mathbf{z}_i y_i \\ \hat{\mathbf{B}}_r &= \left( \sum_{i \in s} w_i a_i c_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in s} w_i a_i c_i \mathbf{z}_i y_i \end{aligned}$$

# Imputation par la régression modifiée

- Sous l'AMNRG, les valeurs imputées adéquates sont données par

$$y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r$$

- Si  $\hat{p}_i \approx p_i$  l'estimateur ajusté est approximativement sans biais sous l'AMNRG
- Si le modèle d'imputation est correctement spécifié, l'estimateur ajusté est approximativement sans biais sous l'AMI



Valide sous les deux approches

# Caractéristiques

- Si  $\hat{p}_i \approx \hat{p}$ , on a  $\tilde{\mathbf{B}}_r = \hat{\mathbf{B}}_r$
- Ne requiert les variables indicatrices de réponse  $a_i$  ou les probabilités de réponse  $\hat{p}_i$  dans le fichier
- Le poids  $w_i \times \frac{(1 - \hat{p}_i)}{\hat{p}_i}$  est obtenu en haussant le poids de sondage  $w_i$  pour les unités qui ont une faible probabilité de réponse et vice versa
- Généralisation de Brewer (JASA, 1979)
- Les valeurs imputées par la régression modifiée peuvent être obtenues par la technique de l'imputation par calage (Beaumont, 2005)

# Comparaisons

	Approche par modèle de non-réponse uniforme	Approche par modèle d'imputation	Approche par modèle de non-réponse généralisé
$\hat{Y}_I$ avec $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$	Sans biais	Sans biais	Biaisé
$\hat{Y}_I^a$ avec $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r$	Sans biais	Sans biais	Sans biais
$\hat{Y}_I$ avec $y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r$	Sans biais	Sans biais	Sans biais

# Choix optimal de $\beta$

- Au lieu de chercher la valeur de  $\beta$  qui garantit un estimateur approximativement sans biais, on peut chercher la valeur de  $\beta$  qui minimise l'EQM conditionnelle, donnée par

$$EQM(\hat{Y}_I | s) = V_r(\hat{Y}_I | s) + \text{Biais}(\hat{Y}_I | s)^2$$

- Le choix optimal,  $\tilde{\mathbf{B}}_{opt}$ , est relativement complexe mais sous certaines conditions, on peut montrer que

$$\tilde{\mathbf{B}}_{opt} = \tilde{\mathbf{B}} + O\left(\frac{1}{n}\right)$$

→ Le choix  $\tilde{\mathbf{B}}$  est presque optimal pour de grandes tailles d'échantillon

# Modéliser les deux?

- L'imputation par la régression modifiée mène à un estimateur imputé approximativement sans biais si l'un des deux modèles est mal spécifié
- Si le mécanisme de non-réponse est **confondu**, alors l'imputation par la régression modifiée mènera généralement à des estimateurs **ayant un plus petit biais** que ceux obtenus par l'imputation par la régression traditionnelle

$$y_i^* = \mathbf{z}_i' \tilde{\mathbf{B}}_r,$$

$$\tilde{\mathbf{B}}_r = \left( \sum_{i \in S} w_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} a_i c_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in S} w_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} a_i c_i \mathbf{z}_i y_i$$

# Imputation par la régression aléatoire

Imputation par la régression aléatoire

=

Imputation par la régression déterministe + résidu aléatoire ajouté

- Imputation par la régression aléatoire traditionnelle: Résidu pour

l'unité  $i$  est tiré avec remise avec probabilité  $w_i / \sum_{j \in S} w_j a_j$

- Imputation par la régression aléatoire modifiée: Résidu pour

l'unité  $i$  est tiré avec remise avec probabilité  $\tilde{w}_i / \sum_{j \in S} \tilde{w}_j a_j$

$$\tilde{w}_i = w_i \left( \frac{1 - \hat{p}_i}{\hat{p}_i} \right)$$

# Conclusions

- L'imputation modifiée mène à des estimateurs valides sous les deux approches
- Le cas des domaines a été étudiée : Estimateur ajusté quand les domaines ne sont pas connus à l'étape de l'imputation
- Estimation de la variance a été étudiée : Approche renversée de Fay (1991) et la méthode de Binder (1983) pour la linéarisation en série de Taylor.