



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

DAVID HAZIZA & J.N.K. RAO

Statistique Canada & Université Carleton

Montréal

24 Septembre 2004

Plan de la présentation

- **Introduction**
- **L'estimateur imputé**
- **Imputation par le ratio**
- **Mécanisme de non-réponse**
- **Approches pour l'inférence**
- **Imputation par le ratio modifié**
- **Étude par simulation**
- **Estimation de la variance**
- **Inférence pour des domaines**
- **Conclusions**

Niveaux de non-réponse

Non-réponse totale:

- Absence complète d'information sur une unité.

Non-réponse partielle:

- Certaines (mais pas toutes) variables recueillies

Traitement de la non-réponse

Non-réponse totale:

En général, on utilise des **méthodes de repondération** qui consiste à hausser le poids des répondants pour compenser pour les non-répondants

Non-réponse partielle:

En général, on utilise **l'imputation** qui consiste à créer une unique valeur artificielle pour boucher le trou de la valeur manquante

Avantages de l'imputation

1. L'imputation simple mène à la création d'un **fichier de données complet**
2. Les résultats issus de différentes analyses seront vraisemblablement **cohérents**
3. Contrairement aux méthodes de repondération, l'imputation permet l'utilisation d'un **poids d'échantillonnage unique**
4. **L'information disponible** sur les répondants partiels peut être utilisée comme information auxiliaire pour améliorer la qualité des valeurs imputées

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

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

- $w_i = 1/\pi_i$ désigne le poids de sondage de l'unité i
- π_i désigne la probabilité d'inclusion de l'unité i dans l'échantillon s ; $i = 1, \dots, N$.
- Échantillonnage aléatoire simple sans remise: $w_i = N/n$

Estimation: réponse complète

- Sous l'approche traditionnelle en sondages (**approche design-based**), le vecteur $\mathbf{y} = (y_1, \dots, y_N)'$ est traité comme fixe
- Soit δ_i la variable indicatrice de sélection dans l'échantillon

$$\delta_i = \begin{cases} 1 & \text{si } i \in s \\ 0 & \text{sinon} \end{cases}$$

- Cette variable joue un rôle crucial dans le contexte de l'inférence
- L'estimateur de HT **est sans biais sous le plan de sondage**

$$E_p(\hat{Y}_{HT}) = Y,$$

où $E_p(\cdot)$ désigne l'espérance par rapport au plan de sondage

Non-réponse: estimateur imputé

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

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

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 le ratio

- Soit z une **variable auxiliaire** disponible pour toutes les unités échantillonnées
- L'imputation par le ratio **déterministe** utilise les valeurs imputées

$$y_i^* = \frac{\bar{y}_r}{\bar{z}_r} z_i \equiv \hat{R}_r z_i,$$

où $(\bar{y}_r, \bar{z}_r) \equiv \sum_{i \in S} w_i a_i (y_i, z_i) / \sum_{i \in S} w_i a_i$ désignent la moyenne des répondants pour les variables y et z , respectivement.

Remarque: On utilise les poids de sondages dans la construction des valeurs imputées

Imputation par le ratio

- Les valeurs imputées peuvent être obtenues en ajustant le modèle de régression

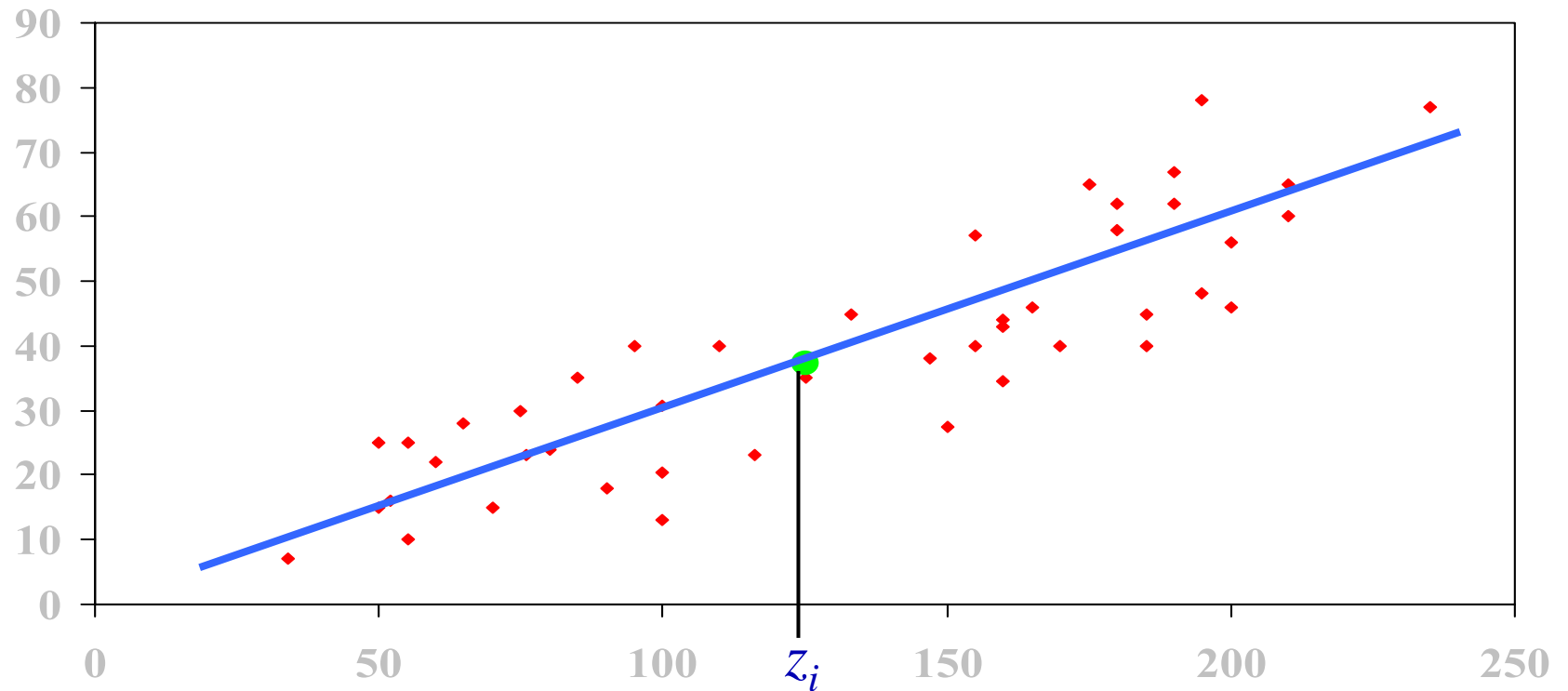
$$m : y_i = \beta z_i + \varepsilon_i,$$
$$E_m(\varepsilon_i) = 0, E_m(\varepsilon_i, \varepsilon_j) = 0, i \neq j,$$

$$V_m(\varepsilon_i) = \sigma^2 z_i,$$

au moyen des unités répondantes.

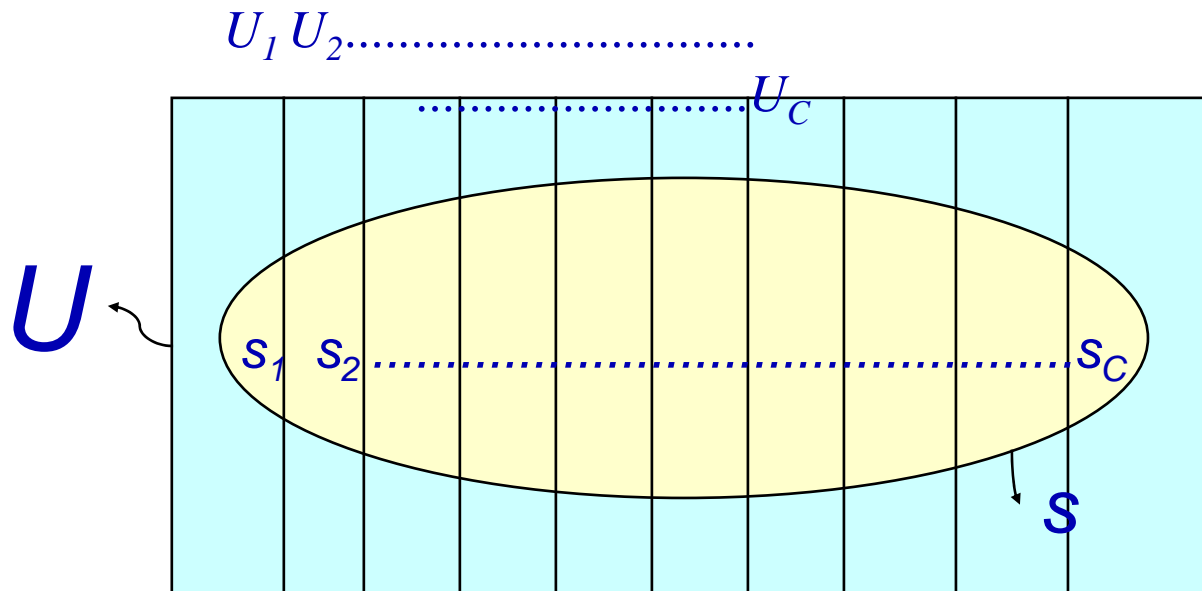
- Les valeurs imputées, $y_i^* = \frac{\bar{y}_r}{\bar{z}_r} z_i$, sont obtenues en effectuant une régression pondérée (avec poids w_i / z_i)

Imputation par le ratio



Classes d'imputation

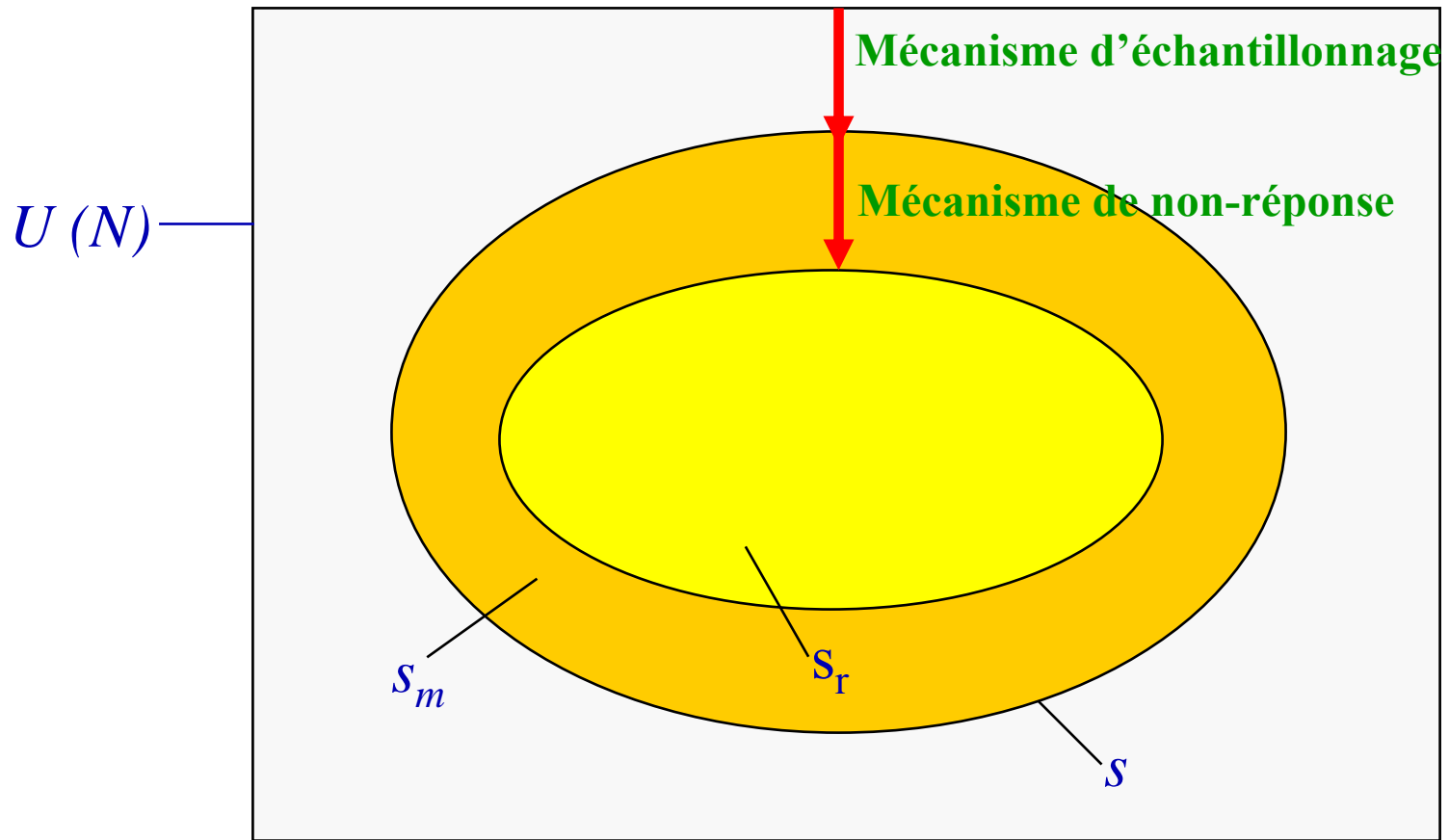
- En pratique, on forme des classes d'imputation et on impute indépendamment dans chaque classe.
- On considère le cas d'une seule classe.
- On suppose que la population U est divisée en classes mutuellement exclusives U_1, U_2, \dots, U_c . Soit $s_c = s \cap U_c$



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**
- Cette distribution n'est généralement pas connue \Rightarrow établir des hypothèses
- Soit $p_i = P(a_i = 1 | s; i \in s)$ la probabilité de réponse pour l'unité $i \Rightarrow$ Les p_i ne sont pas connues.
- On suppose que les unités répondent indépendamment les unes des autres

Mécanisme de non-réponse



Mécanisme de non-réponse

- On distingue 3 types de mécanismes 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)

On suppose que le mécanisme de non-réponse est uniforme, $p_i = p$

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

On suppose que le mécanisme de non-réponse est non-confondu et on fait appel à un modèle d'imputation

$$m : y_i = \beta z_i + \varepsilon_i,$$

$$E_m(\varepsilon_i) = 0, E_m(\varepsilon_i, \varepsilon_j) = 0, i \neq j, V_m(\varepsilon_i) = \sigma^2 z_i$$

Une troisième approche

Approche par Modèle de Non-Réponse Généralisé (AMNRG)

On suppose que 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 disponible pour toutes les unités échantillonnées.

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

Décomposition de l'erreur totale

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

$$\hat{Y}_{I23} - Y = (\hat{Y}_{23} - Y) + (\hat{Y}_{I23} - \hat{Y}_{23})$$

erreur totale erreur due à l'échantillonnage erreur due à la non-réponse

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

→ $\hat{Y} - Y$ est fixe et non-aléatoire

Le biais conditionnel de non-réponse sous l'AMNRU & l'AMNRG

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

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

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

Le biais conditionnel de non-réponse sous l'AMI

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

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

- Sous l'AMI, $E_m(y_i | s, s_r) = \beta x_i$

Le biais: Imputation par le ratio

- L'estimateur imputé est égal à

$$\hat{Y}_I = \frac{\bar{y}_r}{\bar{z}_r} \hat{Z}_{HT}$$

où $\hat{Z}_{HT} = \sum_{i \in s} w_i z_i$.

- **AMNRU**: $\text{Biais}(\hat{Y}_I | s) \approx 0$
- **AMI**: $\text{Biais}(\hat{Y}_I | s) \approx 0$
- **AMNRG**: $\text{Biais}(\hat{Y}_I | s) \neq 0$, en général

Imputation par le ratio modifié

- Les valeurs imputées, $y_i^* = \frac{\bar{y}_r}{\bar{z}_r} z_i$, ne sont donc pas adéquates sous l'AMRG
- Le but est donc 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, $y_i^* = Rz_i$, où R est supposé connu pour l'instant


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


- On détermine R tel que

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

Imputation par le ratio modifié


$$\text{Biais}(\hat{Y}_I | s) = -\sum_{i \in s} w_i (1 - p_i) (y_i - R z_i) = 0 \Leftrightarrow \tilde{R} = \frac{\sum_{i \in s} w_i (1 - p_i) y_i}{\sum_{i \in s} w_i (1 - p_i) z_i}$$

- \tilde{R} 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


$$\hat{R}_r^* = \frac{\sum_{i \in s} w_i a_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} y_i}{\sum_{i \in s} w_i a_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} z_i}$$

Imputation par le ratio modifié

- Sous l'AMNRG, on a $E_r(\hat{R}_r^* | s) \approx \tilde{R}$ si $\hat{p}_i \approx p_i$

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

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

$$y_i^* = \hat{R}_r^* z_i,$$

Remarques

- Sous l'AMNRU, on a $\hat{R}_r^* = \hat{R}_r$ ce qui nous ramène au cas de l'imputation par le ratio 'traditionnelle'
- Le coefficient \hat{R}_r^* est l'estimateur des moindres carrés pondérés obtenu avec les poids

$$\frac{w_i}{z_i} \times \frac{(1 - \hat{p}_i)}{\hat{p}_i}$$

- 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 à réduire celui-ci pour les unités qui ont une grand probabilité de réponse

Remarques

- Dans le cas de l'imputation par le ratio pondéré, l'estimateur imputé \hat{Y}_I est **approximativement sans biais** sous l'AMNRG et l'AMI
- L'estimateur est donc robuste en ce sens qu'il est valide sous les deux approches
- Dans la littérature, l'expression '**double robustness**' est souvent utilisée
- Généralisation de Brewer (JASA, 1979)

Choix optimal de R

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

$$\begin{aligned} EQM(\hat{Y}_I | s) &= V_r(\hat{Y}_I | s) + \text{Biais}(\hat{Y}_I | s)^2 \\ &= \sum_{i \in S} w_i^2 p_i (1 - p_i) (y_i - Rz_i)^2 + \left[\sum_{i \in S} w_i (1 - p_i) (y_i - Rz_i) \right]^2 \end{aligned}$$

Choix optimal de R

- Le choix optimal de R , \tilde{R}_{opt} , est relativement complexe mais sous certaines conditions, on peut montrer

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

- Donc, le choix \tilde{R} est presque optimal pour de grandes tailles d'échantillon

Étude par simulation

- Nous avons généré une population de taille $N=1000$ comprenant 3 variables: une variable d'intérêt y et 2 variables auxiliaires z_1, z_2
- D'abord les variables z_1 et z_2 ont été indépendamment générées à partir d'une distribution exponentielle de moyenne 4 et 30, respectivement.
- La variable y a été générée selon le modèle

$$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i,$$

où les ε_i sont générées à partir d'une loi normale de moyenne 0 et variance σ^2 .

- La valeur de σ^2 a été déterminée de manière à ce que le R^2 du modèle soit approximativement égale à 0.75.

Étude par simulation

- Dans cette population, nous avons tiré $B = 5000$ échantillons aléatoires simples sans remise, de taille $n = 100$.
- Dans chaque échantillon, la non-réponse à la variable y a été générée selon les mécanismes de non-réponse suivants :

Mécanisme non-confondu : la probabilité de réponse p_{1i} pour l'unité i est donnée par

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i}$$

Mécanisme confondu : la probabilité de réponse p_{2i} pour l'unité i est donnée par

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i$$

Étude par simulation

- Les indicateurs de réponse a_{1i} et a_{2i} sont finalement générés à partir d'une distribution de Bernoulli de paramètre p_{1i} et p_{2i} , respectivement.
- Méthodes d'imputation : imputation par la régression traditionnelle et imputation par la régression modifiée.
- Biais relatif: $100 \times [E_{MC}(\hat{Y}_I) - Y] / Y$
- RRMSE: $100 \times \frac{\sqrt{EQM_{MC}(\hat{Y}_I)}}{Y}$

Modèles utilisés

Modèles d'imputation et de non-réponse utilisés :

Modèle pour y	Ordonnée à l'origine	z_1	z_2
$y(1)$	oui	oui	oui
$y(2)$	Oui	Non	Oui
Modèle pour p_i			
$p(1)$	oui	oui	non
$p(2)$	non	oui	non

Résultats : mécanisme non confondu

Modèle	Biais (traditionnel)	Biais (modifié)	RRMSE (traditionnel)	RRMSE (modifié)
$y(1) - p(1)$	0.19	-0.01	1.85	2.33
$y(2) - p(1)$	5.2	0.16	5.60	2.66
$y(1) - p(2)$	0.19	0.05	1.85	1.88

Résultats: mécanisme confondu

Modèle	Biais (traditionnel)	Biais (modifié)	RRMSE (traditionnel)	RRMSE (modifié)
$y(1) - p(1)$	1.84	1.83	2.55	2.54
$y(2) - p(1)$	4.46	1.84	4.89	2.65
$y(1) - p(2)$	1.84	1.84	2.55	2.55

Moyenne d'un domaine

- En pratique, des estimations sont requises pour plusieurs domaines.
- Dans l'Enquête sur la Population Active Canadienne, des estimations du taux de chômage sont requises au niveau provincial et par groupe d'âge-sexe.
- La moyenne d'un domaine peut s'écrire comme

$$\bar{Y}_d = \frac{\sum_{i \in U} d_i y_i}{\sum_{i \in U} d_i},$$

où $d_i = \begin{cases} 1 & \text{si } i \in \text{domaine } d \\ 0 & \text{sinon} \end{cases}$

- On suppose que d_i est connue pour toutes les unités échantillonnées.

Moyenne d'un domaine

- Les domaines ne sont pas toujours connus à l'étape de l'imputation
- Si on n'en tient pas compte pour construire les valeurs imputé, l'estimateur imputé sera généralement biaisé!

Exemple

Nous avons construit une population de taille $N = 11270$ à partir d'un échantillon de l'EPA. La revue hebdomadaire moyen dans la population est 555\$ par semaine.

Age	15-19	20-24	25-29	30-34	35-39
Revenu	139.7	343.6	513.9	587.2	625.6

Age	40-44	45-49	50-59	60-64	65+
Revenu	661.5	704.5	692.4	629.6	549.2

Estimateur imputé

$$\bar{y}_{dI} = \frac{1}{\sum_{i \in S} w_i d_i} \left[\sum_{i \in S} w_i a_i d_i y_i + \sum_{i \in S} w_i (1 - a_i) d_i y_i^* \right]$$

- Si les domaines sont connus à l'étape de l'imputation, on peut utiliser l'imputation par le ratio modifié à l'intérieur du domaine $\Rightarrow y_i^* = \hat{R}_{dr}^* z_i$

$\Rightarrow \bar{y}_{dI}$ est approximativement sans biais pour \bar{Y}_d .

Estimation de la variance

- Traiter les valeurs imputées comme si elles avaient été observées peut mener à une sous-estimation substantielle de la variance de l'estimateur imputé, surtout si le taux de non-réponse est élevé.
- Plusieurs méthodes d'estimation de la variance ont été proposées
 - Jackknife (Rao et Shao, *Biometrika*, 1992)
 - Bootstrap (Shao et Sitter, *JASA*, 1996)
 - Assistée d'un modèle (Särndal, *SM*, 1992)
 - Approche renversée (Shao et Steel, *JASA*, 1999)

Estimation de la variance

- Nous avons utilisé la méthode de Shao-Steel.
- Nous avons utilisé l'approche de Binder (1983, ISR) pour la linéarisation par séries de Taylor pour des équations d'estimation.
- Tenir compte que la probabilité de réponse p_i est estimée.

Biais de l'estimateur imputé

- Lorsque que l'on ne tient pas compte des domaines, on a

$$y_1^* = \hat{R}_r^* z_i$$

- Sous l'AMNG, le biais de \bar{y}_{dI} est donné par

$$\text{Biais}(\bar{y}_{dI}|s) = -\frac{1}{\sum_{i \in S} w_i d_i} \left[\sum_{i \in S} w_i (1 - p_i) d_i (y_i - \tilde{R} z_i) \right]$$

- On estime le biais par

$$\hat{B}(\bar{y}_{dI}|s) = -\frac{1}{\sum_{i \in S} w_i d_i} \left[\sum_{i \in S} w_i a_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} d_i (y_i - \hat{R}_r^* z_i) \right]$$

Estimateur ajusté

- Un estimateur ajusté est obtenu comme suit :

$$\begin{aligned}\bar{y}_{dI}^a &= \bar{y}_{dI} - \hat{B}(\bar{y}_{dI}|s) \\ &= \frac{1}{\sum_{i \in S} w_i d_i} \left[\sum_{i \in S} \frac{w_i}{\hat{p}_i} a_i d_i (y_i - \hat{R}_r^* z_i) + \sum_s w_i d_i \hat{R}_r^* z_i \right]\end{aligned}$$

- L'estimateur ajusté est sans biais sous L'AMNRG et sous l'AMI
⇒ Robuste.

Remarques

- Lorsque le modèle de non-réponse ne contient que l'ordonnée à l'origine, $\hat{p}_i = \hat{p}$, l'estimateur ajusté devient :

$$\bar{y}_{dI}^a = \hat{p}^{-1} \bar{y}_{dI} + \left(1 - \hat{p}^{-1}\right) \frac{\bar{z}_d}{\bar{z}} \bar{y}_I,$$

Haziza et Rao (2004)

- L'estimateur ajusté coïncide avec l'estimateur obtenu par calage (Beaumont, 2004).
 - Dans le cas du calage, on doit connaître les domaine à l'étape de l'imputation.
 - L'estimateur ajusté requière cependant que les probabilités estimées \hat{p}_i et les indicateurs de réponse soient fournis dans le fichier.

Conclusions

- L'imputation modifiée est robuste.
- Nous avons considéré le cas de l'imputation par la régression déterministe et l'imputation par la régression aléatoire.
- Généraliser ces résultats dans le cas de paramètres bivariés (exemple: coefficient de corrélation).