Multiple robustesse pour le traitement de données manquantes dans les enquêtes

David Haziza

Département de mathématiques et de statistique
Université de Montréal

en collaboration avec
Sixia Chen
WESTAT

Les Journées de Méthodologie Statistique
Paris, France

1 avril 2015

PLAN DE LA PRÉSENTATION

- Double robustesse
- Multiple robustesse
- La méthode proposée
- Estimation de la variance et intervalles de confiance
- Étude par simulation
- Extensions

Cadre de travail

- U: population finie de taille N
- But: estimer le total dans la population U de la variable d'intérêt y,

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

- s: échantillon de taille n tiré selon le plan de sondage $p(s)$
- Absence de non-réponse: estimateur par dilatation

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

  - $w_i = 1/\pi_i$: poids de sondage de l'unité i
  - π_i: probabilité d'inclusion dans l'échantillon pour l'unité i

Cadre de travail

- Non-réponse à la variable y: certaines valeurs de y sont manquantes
- Estimateur imputé:

$$\widehat{Y}_I = \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) y_i^*$$

  - $r_i = 1$ si l'unité i répond à la variable y et $r_i = 0$, sinon.
  - y_i^*: valeur imputée utilisée pour remplacer la valeur manquante y_i

- s_r: ensemble des répondants à la variable y
- s_m: ensemble des non-répondants à la variable y

Imputation déterministe

- $\mathbf{x}$: vecteur de variables auxiliaires disponibles pour toutes les unités échantillonnées (répondants et non-répondants)
- Modèle d'imputation

$$y_i = m(\mathbf{x}_i; \boldsymbol{\beta}) + \epsilon_i,$$

où $\boldsymbol{\beta}$ est un vecteur de coefficients (inconnus).

- Estimateur de $\boldsymbol{\beta}$: solution de l'équation estimante

$$\sum_{i \in s} \phi_i r_i \left\{ y_i - m\left(\mathbf{x}_i; \boldsymbol{\beta}\right) \right\} \frac{\partial m(\mathbf{x}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \quad \longrightarrow \widehat{\mathbf{B}}_r$$

où ϕ_i est un coefficient associé à l'unité i.

- Valeur imputée:

$$y_i^* = m(\mathbf{x}_i; \widehat{\mathbf{B}}_r), \quad i \in s_m$$

- Si le modèle d'imputation est bien spécifié, alors l'estimateur $\widehat{Y}_I$ est asymptotiquement sans biais pour Y quelque soit le choix de ϕ_i.

Imputation par la régression linéaire

- Cas particulier:
  - Imputation par la régression linéaire:

$$m(\mathbf{x}_i; \boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

  - Estimateur de $\boldsymbol{\beta}$:

$$\widehat{\mathbf{B}}_r = \left(\sum_{i \in s} \phi_i r_i \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i \in s} \phi_i r_i \mathbf{x}_i y_i$$

  - Valeur imputée:

$$y_i^* = \mathbf{x}_i^\top \widehat{\mathbf{B}}_r, \quad i \in s_m$$

- Choix usuels de ϕ_i:
  - $\phi_i = 1$: imputation par la régression déterministe non-pondérée
  - $\phi_i = w_i$: imputation par la régression déterministe pondérée
- Avec ces choix de ϕ_i, l'estimateur imputé $\widehat{Y}_I$ est généralement biaisé si le modèle d'imputation est mal spécifié

Double robustesse

- Pour se protéger d'une mauvaise spécification du modèle d'imputation, on peut utiliser un choix alternatif de ϕ_i:

$$\phi_i = \widehat{p_i}^{-1} - 1,$$

où $\widehat{p}_i$ est une estimation de la probabilité de réponse à la variable y pour l'unité i obtenue en ajustant le modèle de non-réponse

$$p_i = p(\mathbf{x}_i; \boldsymbol{\alpha}),$$

où $\boldsymbol{\alpha}$ est un vecteur de coefficients inconnus.

- Ce choix de ϕ_i garantit que l'estimateur imputé $\widehat{Y}_I$ est convergent si le modèle d'imputation et/ou le modèle de non-réponse est correctement spécifié $\longrightarrow$ Double robustesse
- Par exemple, Bang et Robins (2005), Haziza et Rao (2006) et Kim et Haziza (2014).

Multiple robustesse

- En pratique, il pourrait être attrayant d'ajuster plusieurs modèles d'imputation et plusieurs modèles de non-réponse.
- Chacun des modèles peut utiliser un ensemble de prédicteurs différents et/ou une fonction de lien différente.
- Une procédure d'imputation est dite multiple robuste si l'estimateur imputé résultant est convergent si tous les modèles sauf un sont mal spécifiés.
- Concept introduit par Han and Wang (2013)
- Peut être vue comme une extension du concept de double robustesse

Multiple robustesse

3 modèles d'imputation

$$m^1 : y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$m^2 : y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

$$m^3 : y_i = \exp\left(\beta_0 + \beta_1 x_i\right)$$

3 modèles de non-réponse

$$p^1 : p_i = \left\{1 + \exp(\alpha_0 + \alpha_1 x_i)\right\}^{-1}$$

$$p^2 : p_i = \left\{1 + \exp(\alpha_0 + \alpha_1 x_i^2)\right\}^{-1}$$

$$p^3 : p_i = (\alpha_0 + \alpha_1 x_i)^{-1}$$

Multiple robustesse

- Deux classes de modèle:
  - Modèles de non-réponse:

$$\mathcal{C}_1 = \{p^j(\mathbf{x}_i; \boldsymbol{\alpha}^j); j = 1, \ldots, J\}$$

  - Modèles d'imputation:

$$\mathcal{C}_2 = \{m^k(\mathbf{x}_i; \boldsymbol{\beta}^k); k = 1, \ldots, K\}$$

- On ajuste chacun de ces $K + J$ modèles pour obtenir

$$\widehat{\boldsymbol{\alpha}}^1, \ldots \widehat{\boldsymbol{\alpha}}^J \longrightarrow p^1(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^1), \ldots, p^J(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^J)$$

et

$$\widehat{\boldsymbol{\beta}}^1, \ldots, \widehat{\boldsymbol{\beta}}^K \longrightarrow m^1(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^1), \ldots, m^K(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^K)$$

- Estimateurs usuels: moindre carrés, maximum de vraisemblance etc.

Multiple robustesse: méthode proposée

Méthode proposée: deux étapes

(1) Étape de calage: $w_i = 1/\pi_i \longrightarrow \widetilde{w}_i$.

(2) Implémentation: imputation de type régression linéaire

Première étape: calage

- $J + K + 1$ équations de calage:

$$\sum_{i \in s_r} \widetilde{w}_i = \sum_{i \in s} w_i;$$

$$\frac{\sum_{i \in s_r} \widetilde{w}_i \times \left(\frac{1}{p^j(\mathbf{x}_i; \widehat{\alpha}^j)} \right)}{\sum_{i \in s_r} \widetilde{w}_i} = \frac{\sum_{i \in s} w_i \times \left(\frac{1}{p^j(\mathbf{x}_i; \widehat{\alpha}^j)} \right)}{\sum_{i \in s} w_i} \equiv \widehat{L}^j, j = 1, \ldots J;$$

$$\frac{\sum_{i \in s_r} \widetilde{w}_i m^k(\mathbf{x}_i; \widehat{\beta}^k)}{\sum_{i \in s_r} \widetilde{w}_i} = \frac{\sum_{i \in s} w_i m^k(\mathbf{x}_i; \widehat{\beta}^k)}{\sum_{i \in s} w_i} \equiv \widehat{m}^k, k = 1, \ldots K;$$

- Similaire à des équations de calage dans un contexte de *Model Calibration* (Wu et Sitter, 2001).

Première étape: calage

- Méthode linéaire: on cherche des poids calés $\widetilde{w}_i$ tels que

$$\sum_{i \in s_r} (\widetilde{w}_i / w_i - 1)^2$$

est minimum tout en respectant les $J + K + 1$ équations de calage.

- On peut utiliser n'importe quelle fonction de distance; voir Deville and Särndal (1992).
- Poids calés:

$$\widetilde{w}_i = w_i(1 + \widehat{\boldsymbol{\lambda}}_r^\top \mathbf{h}_i) \equiv w_i \widehat{F}_i,$$

où

$$\mathbf{h}_i = \left(1, (1/\widehat{p}_i^1) - \widehat{L}^1, \ldots, (1/\widehat{p}_i^J) - \widehat{L}^J, \widehat{m}_i^1 - \widehat{m}^1, \ldots, \widehat{m}_i^K - \widehat{m}^K\right)^\top$$

et

$$\widehat{\boldsymbol{\lambda}}_r = \left(\sum_{i \in s_r} w_i \mathbf{h}_i \mathbf{h}_i^\top\right)^{-1} \left(\sum_{i \in s} w_i \mathbf{h}_i - \sum_{i \in s_r} w_i \mathbf{h}_i\right).$$

Deuxième étape: imputation par la régression

- Procédure d'imputation multiple robuste:

$$y_i^* = \mathbf{h}_i^\top \widehat{\gamma}_p, \quad i \in s_m,$$

où

$$\widehat{\gamma}_p = \left\{ \sum_{i \in s} w_i r_i \left(\widehat{F}_i - 1 \right) \mathbf{h}_i \mathbf{h}_i^\top \right\}^{-1} \sum_{i \in s} w_i r_i \left(\widehat{F}_i - 1 \right) \mathbf{h}_i y_i.$$

- On obtient facilement ces valeurs imputées par une régression linéaire pondérée (au moyen des poids $w_i(\widehat{F}_i - 1)$) avec comme variable dépendante la variable y et comme prédicteurs le vecteur $\mathbf{h}$
- Estimateur imputé:

$$\widehat{Y}_{MR} = \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i \left(1 - r_i\right) \mathbf{h}_i^\top \widehat{\gamma}_p.$$

Propriétés théoriques

Théorème

Si l'un des modèles d'imputation est correctement spécifié, alors l'estimateur imputé $\widehat{Y}_{MR}$ est convergent.

Théorème

Si l'un des modèles de non-réponse est correctement spécifié, alors l'estimateur imputé $\widehat{Y}_{MR}$ est convergent.

Conclusion: $\widehat{Y}_{MR}$ est multiple robuste

Estimation de la variance: multiple robustesse

- On a obtenu un estimateur ponctuel multiple robuste
- Estimateur de la variance multiple robuste: estimateur convergent pour la variance totale (variance due à l'échantillonnage + variance due à la non-réponse) si tous les modèles sauf un sont mal spécifiés
- Intervalle de confiance de niveau 95% pour Y:

$$\widehat{Y}_{MR} \pm 1.96\sqrt{\widehat{V}_{MR}}.$$

- Taux de couverture: proche de 95% si tous les modèles sauf un sont mal spécifiés

Étude par simulation

- On a généré 4 types de populations (de taille $N = 10,000$).
- Chacune comportait 3 variables: deux variables auxiliaires x et z et une variable d'intérêt y.
- Les valeurs de x ont été générées à partir d'une distribution uniforme.
- Les valeurs de z ont été générées à partir d'une distribution du chi-deux.
- Les valeurs de y ont été générées selon deux modèles

$$\text{IM1: } y_i = 1 + 2x + 3x^2 + \epsilon_i \quad \text{et} \quad \text{IM 2: } y_i = 1 + 2x + 3\exp(x) + \epsilon_i$$

- Dans chaque population, on a assigné une probabilité de réponse selon deux mécanismes de non-réponse

$$\text{NM1: } p_i = \left\{1 + \exp(0.8 + 0.5x_i - 0.3x_i^2)\right\}^{-1}$$

et

$$\text{NM 2: } p_i = 1 - \exp\left[-\exp\left\{0.5 + 0.5x_i - 0.3\exp(x_i)\right\}\right]$$

Étude par simulation

- Quatre types de populations: (IM1-NM1), (IM1-NM2), (IM2-NM1) et (IM2-NM2)
- Pour chaque type de population, on a utilisé 2000 itérations
- Dans chaque population on a tiré des échantillons de taille espérée 300 selon un plan de Poisson avec probabilités d'inclusion proportionnelles à la taille z. On a donc

$$\pi_i = 300 \times \frac{z_i}{\sum_{i \in U} z_i}.$$

- Objectif: estimer la moyenne de la population

$$\overline{Y} = N^{-1} \sum_{i \in U} y_i.$$

Multiple robustesse

2 modèles d'imputation

IM1: $y_i = 1 + 2x + 3x^2 + \epsilon_i$

IM 2: $y_i = 1 + 2x + 3\exp(x) + \epsilon_i$

2 modèles de non-réponse

NM1: $p_i = \left\{1 + \exp(0.8 + 0.5x_i - 0.3x_i^2)\right\}^{-1}$

NM 2: $p_i = 1 - \exp\left[-\exp\left\{0.5 + 0.5x_i - 0.3\exp(x_i)\right\}\right]$

On a calculé 3 estimateurs de $\overline{Y}$:

1. Données complètes (COM): $\widehat{\overline{Y}}_{COM} = \sum_{i\in s} w_i y_i / \sum_{i\in s} w_i$
2. 4 estimateurs doublement robustes: $\widehat{\overline{Y}}_{DR}(1010)$, $\widehat{\overline{Y}}_{DR}(1001)$, $\widehat{\overline{Y}}_{DR}(0110)$ et $\widehat{\overline{Y}}_{DR}(0101)$.
3. 5 estimateurs multiple robustes: $\widehat{\overline{Y}}_{MR}(1110)$, $\widehat{\overline{Y}}_{MR}(1101)$, $\widehat{\overline{Y}}_{MR}(1011)$, $\widehat{\overline{Y}}_{MR}(0111)$ et $\widehat{\overline{Y}}_{MR}(1111)$.

Étude par simulation: Mesures Monte Carlo:

- Biais Monte Carlo:

$$B_{MC}(\widehat{\theta}) = E_{MC}(\widehat{\theta}) - \theta.$$

- Écart-type:

$$SE_{MC}(\widehat{\theta}) = \left[\frac{1}{R}\sum_{r=1}^{R}\left\{\widehat{\theta}_{(r)} - E_{MC}(\widehat{\theta})\right\}^2\right]^{1/2}.$$

- Racine carrée de l'erreur quadratique moyenne:

$$RMSE_{MC}(\widehat{\theta}) = \left\{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_{(r)} - \theta\right)^2\right\}^{1/2}.$$

Étude par simulation: Résultats

Estimateurs	Biais	SE	RMSE	Biais	SE	RMSE
	Scénario: (IM1-NM1)			Scénario: (IM1-NM2)		
$\widehat{\overline{Y}}_{COM}$	-0.0026	0.4347	0.4348	0.0022	0.7540	0.7540
$\widehat{\overline{Y}}_{DR}(1010)$	0.0027	0.4960	0.4960	-0.0057	0.8044	0.8044
$\widehat{\overline{Y}}_{DR}(1001)$	0.0165	0.5158	0.5161	0.0074	0.7946	0.7946
$\widehat{\overline{Y}}_{DR}(0110)$	0.0042	0.5016	0.5016	-0.1150	0.8025	0.8107
$\widehat{\overline{Y}}_{DR}(0101)$	0.1582	0.5293	0.5525	0.0089	0.7995	0.7995
$\widehat{\overline{Y}}_{MR}(1110)$	0.0032	0.4963	0.4963	-0.0002	0.7973	0.7973
$\widehat{\overline{Y}}_{MR}(1101)$	0.0147	0.5028	0.5030	0.0072	0.7948	0.7948
$\widehat{\overline{Y}}_{MR}(1011)$	0.0025	0.4966	0.4966	0.0073	0.7948	0.7948
$\widehat{\overline{Y}}_{MR}(0111)$	0.0032	0.4977	0.4978	0.0079	0.7955	0.7955
$\widehat{\overline{Y}}_{MR}(1111)$	0.0026	0.4967	0.4967	0.0073	0.7945	0.7945

Étude par simulation: Résultats

Estimateurs	Biais	SE	RMSE	Biais	SE	RMSE
	Scénario: (IM2-NM1)			Scénario: (IM2-NM2)		
$\widehat{\overline{Y}}_{COM}$	-0.0026	0.4347	0.4348	0.0022	0.7540	0.7540
$\widehat{\overline{Y}}_{DR}(1010)$	-0.0128	0.4959	0.4961	-0.1998	0.7981	0.8228
$\widehat{\overline{Y}}_{DR}(1001)$	0.1990	0.5196	0.5564	-0.0079	0.7958	0.7959
$\widehat{\overline{Y}}_{DR}(0110)$	-0.0155	0.5054	0.5057	-0.0109	0.8152	0.8153
$\widehat{\overline{Y}}_{DR}(0101)$	-0.0155	0.5337	0.5339	-0.0106	0.8034	0.8035
$\widehat{\overline{Y}}_{MR}(1110)$	-0.0141	0.5091	0.5093	0.0318	0.8114	0.8121
$\widehat{\overline{Y}}_{MR}(1101)$	-0.0833	0.5550	0.5613	-0.0100	0.8048	0.8049
$\widehat{\overline{Y}}_{MR}(1011)$	-0.0145	0.5021	0.5023	-0.0098	0.7998	0.7999
$\widehat{\overline{Y}}_{MR}(0111)$	-0.0136	0.5075	0.5077	-0.0104	0.8027	0.8027
$\widehat{\overline{Y}}_{MR}(1111)$	-0.0140	0.5167	0.5169	-0.0099	0.8113	0.8114

Remarques finales

- On a introduit le concept d'inférence multiple robuste
- Procédure d'imputation proposée: imputation de type déterministe:

$$y_i^* = \mathbf{h}_i^\top \widehat{\gamma}_p$$

- Dans l'article: imputation aléatoire et imputation fractionnelle multiple robuste
- Extensions actuellement étudiées
  - Inférence multiple robuste pour le traitement de la sous-couverture
  - Inférence multiple robuste pour des quantiles
  - Imputation multiple multiple robuste