Cibler les variables d'intérêt importantes à l'étape du traitement de la non-réponse dans les enquêtes

David Haziza

Département de mathématiques et de statistique
Université de Montréal

Travail en collaboration avec

Sixia Chen (University of Oklahoma)

et

Yimeng Gao (Université de Montréal)

Journées de Méthodologie Statistique 2018
Paris, France

12 juin 2018

Plan

Plan

Pondération dans les enquêtes: les étapes

- Population: $U = \{1, \cdots, i, \cdots, N\}$.
- $y_1, \ldots, y_p$: p variables d'intérêt
- y_{ji}: valeur de la variable y_j pour l'unité i, $i = 1, \cdots, N$.
- Objectif: Estimer les p totaux

$$t_{y_j} = \sum_{i \in U} y_{ji}, \quad j = 1, \cdots, p.$$

- Estimateurs:

$$\widehat{t}_{y_j,w} = \sum_{i \in S_r} w_i y_{ji}, \quad j = 1, \cdots, p,$$

où w_i désigne le poids final associé à l'unité i.

Pondération dans les enquêtes: les étapes

- Processus de pondération usuel:

$$\underbrace{d_i = \pi_i^{-1}}_{\text{poids de base}} \longrightarrow \underbrace{\widetilde{w}_i = d_i/\widehat{p}_i}_{\text{Poids ajustés pour la non-réponse}} \longrightarrow \underbrace{w_i = \widetilde{w}_i \times f_i}_{\text{Poids finaux (calés)}}$$

- Correction de la non-réponse: réduire le biais de non-réponse
- Calage: garantir la cohérence entre les estimations issues de l'enquête et des totaux connus au niveau de la population
- On focalise sur l'étape de correction de la non-réponse

Pondération par l'inverse de la probabilité de réponse

- On postule un modèle de non-réponse qui est un ensemble d'hypothèses à propos du mécanisme (inconnu) de non-réponse
- Supposons que la probabilité de réponse p_i est liée à un vecteur de variables complètement observées $\mathbf{v}_i$:

$$p_i = f(\mathbf{v}_i),$$

  pour une certaine fonction inconnue $f(\cdot)$.
- Afin d'estimer p_i, on peut avoir recours à
  - des méthodes paramètriques (par exemple, la régression logistique).
  - des méthodes non-paramétriques (méthode des scores, arbres de régression, etc)

Pondération par l'inverse de la probabilité de réponse

- Système de pondération ajusté pour la non-réponse:

$$\{\widetilde{w}_i; i \in S_r\},$$

où

$$\widetilde{w}_i = d_i \times \frac{1}{\widehat{p}_i}.$$

- En appliquant ce système de pondération aux variables $y_1, \cdots, y_p$, on obtient les p estimateurs ajustés pour la non-réponse (propensity score adjusted estimators)

$$\widehat{t}_{y_j,\widetilde{w}} = \sum_{i \in S_r} \widetilde{w}_i y_{ji}, \quad j = 1, \cdots, p.$$

- Les estimateurs $\widehat{t}_{y_1,\widetilde{w}}, \cdots, \widehat{t}_{y_p,\widetilde{w}}$ exhiberont un biais négligeable si le modèle de non-réponse est correctement spécifié

Pondération par l'inverse de la probabilité de réponse

- Le travail de modélisation est important et comprend deux étapes principales:
  - La sélection d'un vecteur de variables $\mathbf{v}_i$ explicatives de la non-réponse ET liées aux variables d'intérêt $y_1, \cdots, y_p$.
  - Le choix d'un modèle décrivant la relation entre la variable indicatrice de réponse r_i et le vecteur de variables $\mathbf{v}_i$.
- L'utilisation de variables $\mathbf{v}_i$ hautement prédictives de r_i tend à conduire à des probabilités de réponse $\widehat{p}_i$ très petites $\longrightarrow$ les poids $\widetilde{w}_i$ seront potentiellement très dispersés.
- Si les poids $\widetilde{w}$ sont peu ou pas liées aux variables $y_1, \cdots, y_p$, les estimateurs résultants seront potentiellement inefficaces (grande variance).

Illustration

- Si une variable v est liée à la probabilité de réponse mais n'est pas liée aux variables d'enquête $y_1, \cdots, y_p$, il ne faut pas inclure v dans le modèle de non-réponse $\longrightarrow$ aucun impact sur le biais mais rendra les estimateurs potentiellement inefficaces
- On a généré une population de taille $N = 5000$ avec 7 variables: une variable d'intérêt y et 6 variables auxiliaires v_1-v_6.
- Modèle pour y:

$$y_i = \beta_0 + \beta_1 v_{1i} + \beta_2 v_{2i} + \epsilon_i,$$

où les ϵ_i ont été générées à partir d'une $\mathcal{N}(0, \sigma^2)$.

- $R^2 \approx 60\%$
- De la population, on a tiré $10,000$ échantillons, de taille $n = 200$, selon un plan aléatoire simple sans remise.

Illustration

- Dans chaque échantillon, on a assigné à chaque unité une probabilité de réponse p_i selon

$$p_i = \{1 + \exp(\gamma_0 + \gamma_1 v_{1i} + \cdots + \gamma_6 v_{6i})\}^{-1}.$$

- Taux de réponse global: environ 70%.
- On a généré la non-réponse: $r_i \sim \mathcal{B}(1, p_i)$.
- Objectif: estimer $t_y = \sum_{i \in U} y_i$.
- Note: Les variables v_1-v_6 sont observées pour toutes les unités dans l'échantillon (répondants et non-répondants).

Illustration

- On a calculé deux estimateurs de t_y:
  - L'estimateur non-ajusté: $\widehat{t}_{y,resp} = N\widehat{\overline{Y}}_r$;
  - L'estimateur ajusté:

$$\widehat{t}_{y,\widetilde{w}} = \sum_{i \in S_r} \frac{d_i}{\widehat{p}_i} y_i = \sum_{i \in S_r} \widetilde{w}_i y_i,$$

    où $\widehat{p}_i$ est obtenue au moyen d'une régression logistique.
- Mesures Monte Carlo:
  - Biais relatif:

$$RB_{MC}(\widehat{t}) = \frac{1}{10,000} \sum_{k=1}^{10,000} \frac{(\widehat{t}_{(k)} - t_y)}{t_y} \times 100.$$

  - Erreur quadratique moyenne:

$$MSE_{MC}(\widehat{t}) = \frac{1}{10,000} \sum_{k=1}^{10,000} \left(\widehat{t}_{(k)} - t_y\right)^2.$$

Illustration

Estimateur	$\widehat{t}_{y,resp}$	$\widehat{t}_{y,\widetilde{w}}$ v_1	$\widehat{t}_{y,\widetilde{w}}$ v_1-v_2	$\widehat{t}_{y,\widetilde{w}}$ v_1-v_3	$\widehat{t}_{y,\widetilde{w}}$ v_1-v_4	$\widehat{t}_{y,\widetilde{w}}$ v_1-v_5	$\widehat{t}_{y,\widetilde{w}}$ v_1-v_6
Biais relatif en (%)	-3.2	-2.3	-0.1	-0.1	-0.1	-0.1	-0.1
EQM	7.75	4.54	1.66	1.69	1.78	1.94	2.76
$S^2_{\widetilde{w}}$	0	212	1527	2093	3465	5250	1×10^6

Note: $S^2_{\widetilde{w}}$ désigne la variabilité des poids ajustés pour la non-réponse.

Plan

Méthode proposée

- Conclusion: le vecteur **v** ne devrait contenir que les variables liées à la fois à la probabilité de réponse et aux variables d'intérêt.
- Dans le cas d'une enquête qui recueille un grand nombre de variables d'intérêt, choisir le vecteur **v** peut être ardu.
- Souvent, on est en mesure d'identifier un petit nombre de variables d'intérêt jugées importantes (disons entre une et cinq).
- Sans perte de généralité, supposons que les G premières variables d'intérêt sont jugées importantes.
- On suppose que la variable $y_j, j = 1, \cdots, G$ obéit au modèle suivant:

$$y_{ji} = m^{(j)}(\mathbf{z}_i^{(j)}; \boldsymbol{\beta}^{(j)}) + \epsilon_i^{(j)}, \quad j = 1 \cdots, G,$$

où

  - $m^{(j)}(\cdot; \boldsymbol{\beta}^{(j)})$: fonction inconnue;
  - $\mathbf{z}^{(j)}$: vecteur de variables auxiliaires associé à la variable y_j.

Méthode proposée

- Note: différentes fonctions de lien et différentes variables auxiliaires pour chaque variable d'intérêt.
- Pour chaque modèle, on obtient les estimateurs de $\boldsymbol{\beta}^{(j)}$, $j = 1, \cdots, G$ $\longrightarrow$ Estimateurs du maximum de vraisemblance, moindres carrés, etc.
- Pour l'unité $i \in S$ (répondantes et non-répondantes), on obtient les G valeurs prédites

$$m^{(1)}(\mathbf{z}_i^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)}), \cdots, m^{(G)}(\mathbf{z}_i^{(G)}; \widehat{\boldsymbol{\beta}}^{(G)}).$$

- Au total, pour chaque $i \in S$, on dispose de

$$\widehat{p}_i, m^{(1)}(\mathbf{z}_i^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)}), \cdots, m^{(G)}(\mathbf{z}_i^{(G)}; \widehat{\boldsymbol{\beta}}^{(G)}).$$

Méthode proposée

- On cherche à obtenir des poids de calage $\widetilde{w}_i$ aussi proches que possible des poids initiaux d_i tels que les $G+2$ contraintes suivantes sont satisfaites:

$$\sum_{i \in S_r} \widetilde{w}_i = \sum_{i \in S} d_i,$$

$$\sum_{i \in S_r} \widetilde{w}_i L\{1/\widehat{p}_i\} = \sum_{i \in S} d_i L\{1/\widehat{p}_i\}$$

et

$$\sum_{i \in S_r} \widetilde{w}_i m^{(j)}(\mathbf{z}_i^{(j)}; \widehat{\boldsymbol{\beta}}^{(j)}) = \sum_{i \in S} d_i m^{(j)}(\mathbf{z}_i^{(j)}; \widehat{\boldsymbol{\beta}}^{(j)}) \quad j = 1, \ldots, G.$$

- La fonction $L(t)$ désigne l'inverse de la fonction de calage $F(\cdot)$ (voir ci-dessous).

Méthode proposée

- On cherche des poids calés $\widetilde{w}_i$ tels que

$$\sum_{i \in S_r} G(\widetilde{w}_i / d_i)$$

est minimum tout en satisfaisant les $G + 2$ contraintes.

- Les poids $\widetilde{w}_i$ sont donnés par

$$\widetilde{w}_i = d_i \times F(\widehat{\boldsymbol{\lambda}}_r^\top \mathbf{h}_i),$$

où

  - $F(.)$ désigne la fonction de calage;
  - $\widehat{\boldsymbol{\lambda}}_r^\top$ est un vecteur de taille $G + 2$ de coefficients estimés et

$$\mathbf{h}_i = \left(1, \widehat{L}_i - \widehat{\overline{L}}, \widehat{m}_i^{(1)} - \widehat{\overline{m}}^{(1)}, \ldots, \widehat{m}_i^{(G)} - \widehat{\overline{m}}^{(G)}\right)^\top.$$

Méthode proposée

- En appliquant le système de pondération $\{\widetilde{w}_i; i \in S_r\}$ à la variable y_j, on arrive à

$$\widehat{t}_{y_j,\widetilde{w}} = \sum_{i \in S_r} \widetilde{w}_i y_{ji}, \quad j = 1, \cdots, p.$$

Theorem

Si le modèle de non-réponse $p(\mathbf{v}_i)$ est correctement spécifié, alors $\widehat{t}_{y_j,\widetilde{w}}$ est convergent pour t_{y_j}, $j = 1, \cdots, p$.

Theorem

Si le modèle de non-réponse $p(\mathbf{v}_i)$ et/ou le modèle pour y_j est correctement spécifié, alors $\widehat{t}_{y_j,\widetilde{w}}$ est convergent pour t_{y_j}, $j = 1, \cdots, G$.

$\longrightarrow$ *Double robustesse*

Plan

Étude par simulation

- Nous avons généré 1 population de taille $N = 10000$.
  - 3 variables d'intérêt: y_1, y_2 et y_3
  - 8 variables auxiliaires $x_1 - x_8$
  - y_1: variable continue générée à partir d'un modèle de régression linéaire;
  - y_2: variable dichotomique générée à partir d'un modèle de régression logistique;
  - y_3: variable de comptage générée à partir d'un modèle de Poisson;

Variables	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y_1	✓	✓	✓	X	X	X	✓	✓
y_2	✓	✓	✓	X	X	X	✓	✓
y_3	✓	✓	✓	X	X	X	✓	✓
p_i	✓	✓	✓	✓	✓	✓	X	X

Étude par simulation

- $R^2 \approx 0.5$ pour chacun des modèles $(y_1 - y_3)$
- De la population, on a tiré 1000 échantillons selon un plan aléatoire de taille $n = 500$.
- On a généré la non-réponse selon un modèle logistique:

$$p_i = \{1 + \exp(\gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_6 x_{6i})\}^{-1}.$$

- Taux de réponse global: 70% environ
- Objectif: estimer t_{y_1}, t_{y_2} et t_{y_3}

Étude par simulation

- On a calculé deux estimateurs de t_{y_1}, t_{y_2} et t_{y_3} de la forme

$$\widehat{t}_{y_j,\widetilde{w}} = \sum_{i \in S_r} \widetilde{w}_i y_{ji}, \quad j = 1, \cdots, 3.$$

  - L'estimateur ajusté usuel avec $\widetilde{w}_i = d_i/\widehat{p}_i$ et $\widehat{p}_i$ est obtenue au moyen d'une régression logistique avec $x_1 - x_6$ (bon modèle)+ méthode des scores avec 5 classes
  - L'estimateur ajusté proposé avec $\widetilde{w}_i$ obtenue au moyen de la méthode de calage linéaire avec les 5 équations de calage suivantes:

$$\sum_{i \in S_r} \widetilde{w}_i = \sum_{i \in S} d_i, \quad \sum_{i \in S_r} (\widetilde{w}_i/\widehat{p}_i) = \sum_{i \in S} d_i/\widehat{p}_i$$

$$\sum_{i \in S_r} \widetilde{w}_i \widehat{y}_{1i} = \sum_{i \in S} d_i \widehat{y}_{1i}, \quad \sum_{i \in S_r} \widetilde{w}_i \widehat{y}_{2i} = \sum_{i \in S} d_i \widehat{y}_{2i}, \quad \sum_{i \in S_r} \widetilde{w}_i \widehat{y}_{3i} = \sum_{i \in S} d_i \widehat{y}_{3i}.$$

  - Les prédictions $\widehat{y}_{1i}, \widehat{y}_{2i}$ et $\widehat{y}_{3i}$ ont été obtenues au moyen des bons modèles (bonne fonction de de lien + x_1, x_2, x_3, x_7, x_8 comme prédicteurs)

Étude par simulation: Résultats

- Biais relatif négligeable dans tous les cas (pas suprenant)
- Efficacité relative:

$$\text{ER} = 100 \times \frac{EQM(\text{méthode usuelle})}{EQM(\text{méthode proposée})}.$$

Variable	y_1	y_2	y_3
ER (logistique)	120	139	145
ER (méthode des scores)	110	125	135

Remarques finales

- On peut remplacer toutes les méthodes paramétriques (pour obtenir $\widehat{p}_i, \widehat{y}_{1i}, \cdots, \widehat{y}_{Gi}$) par des méthodes non-paramétriques $\longrightarrow$ méthodes des scores, arbres de régression, splines, etc. $\longrightarrow$ Hic: théorie beaucoup plus complexe.
- Estimation de la variance: par séries de Taylor ou par méthode de rééchantillonnage (jackknife)
- Attention: On peut imaginer des situations pour lesquelles la méthode proposée sera moins efficace que la méthode usuelle.
- Comment savoir si la méthode proposée est préférable à la méthode usuelle dans une enquête particulière $\longrightarrow$ comparaison des variances estimées
- D'autres scénarios de simulation sont en cours de test $\rightarrow$ mauvaise sépécifications des modèles, non-réponse non-ignorable, méthodes d'estimation non-paramétriques.