Régression binaire par « capture-recapture »

Jean-Baptiste ANTENORD[1,2] Etienne BILLETTE de VILLEMEUR[2]

[1]Université Quisqueya (CREGED / Haïti)

[2]Université de Lille (LEM, UMR9221 / Lille, France)

24-26 octobre 2018
10ème Colloque Francophone sur les Sondages

Les méthodes de « capture-recapture » (MCR)

De l'étude des populations animales...

Comment estimer la taille d'une population animale ?

- Pas de recensement
- Pas de base de sondage

=> Procéder avec des « captures » répétées

- Les animaux sont marqués et relâchés après chaque capture
- On infère la taille de la population non-observée de l'observation

... à celle des sociétés humaines

Pour des Populations difficiles à enquêter (PDE)
(Enfants de rue, sans-abris, extrême pauvreté...)

- Combinaison de différentes sources d'information
- Sondages aléatoires répétés

De l'étude des populations animales...

Comment estimer la taille d'une population animale ?

- Pas de recensement
- Pas de base de sondage

=> Procéder avec des « captures » répétées

- Les animaux sont marqués et relâchés après chaque capture
- On infère la taille de la population non-observée de l'observation

... à celle des sociétés humaines

Pour des Populations difficiles à enquêter (PDE)
(Enfants de rue, sans-abris, extrême pauvreté...)

- Combinaison de différentes sources d'information
- Sondages aléatoires répétés

Quelle proportion d'individus avec la caractéristique x a également la caractéristique y ?

Objectifs de l'étude :

Proposer un estimateur de probabilité conditionnelle

dans le cas où

- On dispose de ***deux échantillons indépendants***
- ... ***pas*** nécessairement ***représentatifs***.

Principe des Méthodes de « Capture-Recapture » :
Extraire l'information contenue dans K observations partielles, pour en inférer les caractéristiques de la population non observée.

Soit

- $\mathscr{U} = \{1, 2, ..., N\}$, la population
- n^{yx}, le nombre d'individus de caractéristiques $y \in \{0,1\}$ et $x \in \{0,1\}$
- $\omega^i = (\omega_1^i, \omega_2^i, ..., \omega_K^i)$, historique des « captures » de l'individu $i \in \mathscr{U}$
  ($\omega_k^i = 1$ *si l'individu apparaît dans la liste* k *et* $\omega_k^i = 0$ *sinon*)
- n_ω^{yx}, nombre d'individus de caractéristiques yx et d'historique de capture ω

Par définition :

$$n^{yx} = \sum_{\omega \in \Omega} n_\omega^{yx}.$$

Cas de K=2 listes indépendantes

- Ensemble des historiques de capture-recapture : $\Omega = \{11; 10; 01; 00\}$
  - $\omega = 11$: présent dans les deux listes
  - $\omega = 00$: jamais observé
- Pour un individu de caractéristiques y, x :
  - r_1^{yx}, probabilité d'être inclus dans la liste 1
  - r_2^{yx}, probabilité d'être inclus dans la liste 2

	Présent dans la liste 2	Absent de la liste 2
Présent dans la liste 1	$n_{11}^{yx} = n^{yx}\left(r_1^{yx}\right)\left(r_2^{yx}\right)$	$n_{10}^{yx} = n^{yx}\left(r_1^{yx}\right)\left(1 - r_2^{yx}\right)$
Absent de la liste 1	$n_{01}^{yx} = n^{yx}\left(1 - r_1^{yx}\right)\left(r_2^{yx}\right)$	$n_{00}^{yx} = n^{yx}\left(1 - r_1^{yx}\right)\left(1 - r_2^{yx}\right).$

La population non-observée s'élève à

$$n_{00}^{yx} = \frac{n_{01}^{yx}\, n_{10}^{yx}}{n_{11}^{yx}}.$$

Estimateur de probabilité conditionnelle par la MCR

Estimateur de population par la MCR (Rivest et Lavallée, 2012)

$$\widehat{n^{yx}} = n_{01}^{yx} + n_{10}^{yx} + n_{11}^{yx} + \frac{n_{01}^{yx} n_{10}^{yx}}{n_{11}^{yx}} = \frac{1}{n_{11}^{yx}} \left(n_{01}^{yx} + n_{11}^{yx}\right) \left(n_{10}^{yx} + n_{11}^{yx}\right).$$

Estimateurs de probabilité conditionnelles par la MCR :

$$\widehat{\overline{q}} = \mathrm{Prob}\left\{y = 1 \,|\, x = 1\right\} = \frac{\widehat{n^{11}}}{\widehat{n^{11}} + \widehat{n^{01}}}$$

$$= \frac{n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right)}{\left[n_{11}^{01}\left(n_{01}^{11} + n_{11}^{11}\right)\left(n_{10}^{11} + n_{11}^{11}\right) + n_{11}^{11}\left(n_{01}^{01} + n_{11}^{01}\right)\left(n_{10}^{01} + n_{11}^{01}\right)\right]};$$

$$\underline{\widehat{q}} = \mathrm{Prob}\left\{y = 1 \,|\, x = 0\right\} = \frac{\widehat{n^{11}}}{\widehat{n^{11}} + \widehat{n^{01}}} = \ldots$$

Estimation de la variance

(Par la méthode Delta)

$$\widehat{V\left(\widehat{q}\right)} \simeq \frac{\left(\widehat{n^{01}}\right)^2}{\left(\widehat{n^{01}}+\widehat{n^{11}}\right)^4}\widehat{V\left(\widehat{n^{11}_{00}}\right)} + \frac{\left(\widehat{n^{11}}\right)^2}{\left(\widehat{n^{01}}+\widehat{n^{11}}\right)^4}\widehat{V\left(\widehat{n^{01}_{00}}\right)}.$$

La variance asymptotique sur-estime la vraie variance
(Sekar et Deming, 1949 ; Manly, 1969)

$$\left[\frac{\widehat{V\left(\widehat{n^{yx}_{00}}\right)}}{\widehat{V_a\left(\widehat{n^{yx}_{00}}\right)}}\right] < \frac{n^{yx}_{11}}{\widehat{n^{yx}}} = \frac{\left(n^{yx}_{11}\right)^2}{\left(n^{yx}_{01}+n^{yx}_{11}\right)\left(n^{yx}_{10}+n^{yx}_{11}\right)}.$$

Biais de l'estimateur naïf $\widetilde{\overline{q}}$

Estimateur « *naïf* » de $\overline{q} = \text{Prob}\{y = 1 \mid x = 1\}$:

$$\widetilde{\overline{q}} = \frac{o^{11}}{o^{11} + o^{01}},$$

où $o^{yx} = n_{01}^{yx} + n_{10}^{yx} + n_{11}^{yx}$ est l'effectif de caractéristiques yx observé.

- Biais de l'estimateur naïf $\widetilde{\overline{q}}$:

$$b_{\overline{q}} \equiv \widetilde{\overline{q}} - \widehat{\overline{q}} = \widehat{b}\left[\left(\frac{\widehat{n_{00}^{01}}}{o^{01}}\right) - \left(\frac{\widehat{n_{00}^{11}}}{o^{11}}\right)\right],$$

où $\widehat{b} > 0$.

- Condition de biais nul :

$$\frac{\widehat{n_{00}^{11}}}{\widehat{n^{11}}} = \frac{\widehat{n_{00}^{01}}}{\widehat{n^{01}}}.$$

Biais asymptotique de l'estimateur naïf et test de biais nul

Le biais de $\tilde{\tilde{q}}$ est *asymptotiquement nul* si et seulement si les individus de caractéristiques $yx = 11$ et $yx = 01$ ont la même probabilité de ne pas être observés

- *Probabilité qu'un individu de caractéristique yx ne soit* pas *observé :*

$$p_{00}^{yx} = \left(1 - r_1^{yx}\right)\left(1 - r_2^{yx}\right),$$

avec $\widehat{r_1^{yx}} = n_{11}^{yx} / \left(n_{01}^{yx} + n_{11}^{yx}\right)$ et $\widehat{r_2^{yx}} = n_{11}^{yx} / \left(n_{10}^{yx} + n_{11}^{yx}\right)$.

- Asymptotiquement, la condition de biais nul est vérifiée si, au risque α

$$P\left(\left| Z = \frac{\widehat{p_{00}^{11}} - \widehat{p_{00}^{01}}}{\sqrt{\widehat{V\left(\widehat{p_{00}^{11}} - \widehat{p_{00}^{01}}\right)}}} \xrightarrow{\mathscr{L}} N(0,1) \right| \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Conclusion

La contribution

- Nous proposons un estimateur de probabilité conditionnelle
  - Sur la base de deux échantillons *indépendants* mais *non-représentatifs*
  - Pour lequel nous calculons la variance exacte
- Un test statistique est proposé afin d'évaluer la possibilité de se dispenser du redressement statistique sous-jacent à notre estimateur

- Dans de nombreuses circonstances, les populations sont difficiles à rejoindre et on ne dispose pas de bases de sondage
- Par l'application des MCR, il est possible d'obtenir des estimateurs qui s'affranchissent de ces difficultés sur la base d'observations indépendantes et répétées.