"To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling", Roderick J. Little, 2004.

Groupe de lecture - Réunion 3

Marine GUILLERM

INSEE, division MAEE

19 janvier 2015

Introduction

- Les sessions 1 et 2, la question : comment estimer correctement les paramètres d'un modèle avec des données d'enquête ?
- Session 3 : quel apport des modèles dans le cadre d'estimation sur une population finie ?

Sommaire

Notations

- Une population finie de taille N. $Y = (y_1, \ldots, y_N)$ le vecteur des variables.
- But : estimer une quantité $Q(Y)$ sur la population finie (par exemple un total, une moyenne)...
- ...à partir d'un échantillon s. On note Y_{inc} le vecteur des valeurs observées, Y_{exc} son complémentaire.
  $I = (I_1, \ldots, I_N)$ est le vecteur des indicatrices d'inclusion dans l'échantillon.
- On a défini un plan de sondage : à chaque échantillon s est associée une probabilité de sélection $p(s)$.

Les caractéristiques de l'approche sondage

- Les valeurs Y_i sont considérées comme fixes.
- L'aléa provient de la sélection de l'échantillon s.
- L'inférence repose uniquement sur la distribution de I, sur le plan de sondage.

Un estimateur de $Q(Y)$ est $\hat{q}(Y_{inc}, I)$. La variabilité de l'estimateur vient de I, pas des valeurs Y_k qui sont fixes.
Donc, par exemple : $E(\hat{q}) = \sum_s p(s)\hat{q}(Y^s_{inc}, I_s)$.

Les avantages / inconvénients de l'approche sondage

- Produit des estimateurs fiables pour des échantillons de grande taille sous de faibles hypothèses.
  $\rightarrow$ En particulier pas d'hypothèse sur la distribution des Y.
- Asymptotique donc pas forcément adapté pour des petits échantillons.
- Pas de théorie pour construire des estimateurs optimaux.
  $\rightarrow$ Il n'existe pas de plan de sondage et d'estimateurs, qui quelque soit les valeurs de Y, fournissent un estimateur de variance minimum.
- Des calculs de variance compliqués pour les plans de sondage complexes.
- La technique d'inférence n'est plus adaptée quand le plan de sondage est perturbé, par exemples : non réponse, erreurs de mesure.

Sommaire

Model-assisted design-based inference (1/2)

Idée : Utiliser des modèles pour déterminer le plan de sondage, construire/justifier le choix d'un estimateur

... mais l'inférence repose toujours sur le plan de sondage.

Quelques exemples :

- Estimateur de HT
  On considère le modèle $y_i = \beta\pi_i + \pi_i\varepsilon_i$ avec ε_i iid $\sim N(0, \sigma^2)$.
  On trouve $\hat{\beta} = \hat{T}_{HT}/n$.
  $\rightarrow$ HT est un bon estimateur si le modèle décrit bien la population, ie $y_i/\pi_i \sim N(\beta, \sigma^2)$.
- L'estimateur de régression généralisé (GREG)

$$\hat{T}_{GREG} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \in s} \frac{y_i - \hat{y}_i}{\pi_i}$$

avec $\hat{y}_i = x_i\hat{\beta}$, $(x_1, \ldots, x_n)$ un vecteur de variables auxiliaires.

Model-assisted design-based inference (2/2)

- Pseudo-vraisemblance
  - Dans une approche modèle, on estime le paramètre d'intérêt θ en maximisant la (log) vraisemblance.
    On cherche θ tel que :

$$sc_{pop}(\theta) = \sum_{i=1}^{N} \partial \log p(y_i|z_i, \theta)/\partial\theta = 0$$

    Hypothèse : indépendance des observations, on spécifie un modèle.
  - $sc_{pop}(\theta)$ est une quantité sur la population finie, estimée par HT et on résout :

$$sc_{HT}(\theta) = \sum_{i \in s} \frac{1}{\pi_i} \partial \log p(y_i|z_i, \theta)/\partial\theta \qquad (1)$$

    Dans le cas d'une régression linéaire, on trouve l'estimateur OLS pondéré par les π_i.

(1) généralise l'estimateur de HT mais ne résout pas son potentiel manque d'efficacité, par exemple quand un outlier a une proba d'inclusion très faible.

Sommaire

Approche modèle - *Model-based*

- La population totale est un échantillon d'une superpopulation.
- Les valeurs Y_i sont des réalisations de variables aléatoires.
- On spécifie un modèle en choisissant une loi paramétrée pour Y : $p(Y|\theta)$.
- Le modèle est utilisé pour prédire la distribution de Y_{exc}, et donc estimer $Q(Y)$.
- L'inférence repose sur la distribution de Y.

Model-based - approche bayésienne

Ici, présentation plus particulièrement de l'approche bayésienne.
Principe général :
On spécifie :

1. Un modèle paramétrique pour $Y|\theta$.
2. Une distribution a priori de θ.
   $\rightarrow$ Permet d'inclure de l'information a priori sur le paramètre
   ... ou pas $\rightarrow$ on choisit alors un a priori non informatif.

La formule de Bayes permet de trouver la distribution a posteriori Y_{exc}/Y_{inc}, et donc celle de $Q|Y_{inc}$ à partir de laquelle on estime $Q(Y)$.

Exemple : le sondage stratifié (1/3)

- Dans l'approche sondage, l'estimateur stratifié est $\bar{y}_{st} = \sum_{j=1}^{J} \frac{N_j}{N} \bar{y}_j$.
- Approche modèle bayesien
  On spécifie le modèle :

$$y_{ji}|\mu, \sigma^2 \sim N(\mu, \sigma^2)$$
$$\mu, \log \sigma^2 \propto cste$$

  - On trouve $E(\bar{Y}/Y_{inc}) = \bar{y} = \sum_{j=1}^{J} \frac{n_j}{n} \bar{y}_j$.
  - Du point de vue de l'approche sondage, bon estimateur de $\bar{Y}$ si le taux de sondage n_j/N_j le même dans toutes les strates... ce qui est rarement le cas.
  - Le modèle spécifié fait l'hypothèse que la distribution de Y est la même dans chaque strate.
  - Mauvaise spécification du modèle. Il faut tenir compte du plan de sondage.

Sondage stratifié (2/3)

On spécifie le modèle plus élaboré :

$$y_{ji}|\mu_j, \sigma_j^2 \sim N(\mu_j, \sigma_j^2)$$
$$\mu_j, \log \sigma_j^2 \propto cste$$

- Un modèle qui reflète les différences entre strates.
- La sélection est ignorable à l'intérieur de chaque strate.
- On trouve

$$\bar{Y}/Y_{inc}, I, \{\sigma_j^2\} \sim N\left(\sum_{j=1}^{J} \frac{N_j}{N}\bar{y}_j, \sum_{j=1}^{J}\left(\frac{N_j}{N}\right)^2\left(1-\frac{n_j}{N_j}\right)\frac{\sigma_j^2}{n_j}\right)$$

- On retrouve l'estimateur stratifié en sondage.
- Quand on remplace σ_j^2 par leur estimateur s_j^2, on retrouve le même estimateur de variance que dans l'approche sondage, avec le facteur de correction de population finie $1 - n_j/N_j$.

Sondage stratifié (3/3)

Si on trouve la même chose qu'en sondage, quel apport du bayésien ?

- La possibilité d'utiliser la distribution a posteriori de σ_j^2 plutôt que de remplacer σ_j^2 par s_j^2.
  -> important sur de petits échantillons.
- Des modèles encore plus élaborés pour tenir compte d'effets aléatoires strates (modèles multi-niveaux).

Avantages / limites de la modélisation bayesienne

- + semble réconcilier les deux approches : on retrouve les estimateurs de l'approche sondage avec la possibilité de les améliorer
- + produit de bonnes inférences pour des petits échantillons
- - ... mais les choses semblent vite se complexifier
- + on peut inclure de l'information a priori
- + on peut inclure une modélisation pour la non réponse
- - Une mauvaise spécification du modèle produit de mauvais estimateurs
- - Pour bien spécifier le modèle, il faut tenir compte du plan de sondage, pas toujours connu
- +/- les calculs de variance ?