Partie 2
Introduction à l'estimation sur petits domaines

Cette présentation est largement basée sur :

- Ardilly, P. (2005). Panorama des principales méthodes d'estimation sur petits domaines. Actes des Journées de Méthodologie Statistique, Insee.
- Rao, J.N.K (2003). Small Area Estimation. New-York, Wiley.

En résumé

Dans un sondage dans une population finie U, les estimateurs ont généralement un biais nul (ou négligeable), et une variance qui décroît quand la taille d'échantillon augmente. Une enquête de taille raisonnable donnera donc des estimateurs d'une bonne précision.

Pour l'estimation de paramètres portant sur des sous-populations (ou domaines) de U, la taille de l'échantillon exploitable décroît avec la taille du domaine. Si la taille d'échantillon dans le domaine est trop faible, la variance des estimateurs directs peut être prohibitive.

La théorie de l'estimation sur petits domaines vise à mobiliser de l'information externe et/ou de l'information auxiliaire sur le domaine afin de produire des estimateurs alternatifs. Ces estimateurs auront généralement une variance plus faible, mais leur biais dépend de la validité des hypothèses réalisées.

Rappels sur l'estimation en population finie

Estimation

On se place dans le cadre d'une population finie d'individus, notée U. On collecte les valeurs prises par une variable d'intérêt y sur un échantillon S.

L'échantillon S est sélectionné dans U au moyen d'un **plan de sondage** $p(\cdot)$. Le total $t_y = \sum_{k \in U} y_k$ est estimé **sans biais** par l'estimateur de Horvitz-Thompson (ou π-estimateur)

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k \; y_k$$

avec $d_k = 1/\pi_k$ le poids de sondage de l'unité k.

Calcul de précision

Si le plan de sondage $p(\cdot)$ est de taille fixe n, la variance du π-estimateur est donnée par

$$V_p\left[\hat{t}_{y\pi}\right] = -\frac{1}{2}\sum_{k\neq l\in U}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\Delta_{kl}.$$

Dans le cas particulier d'un SRS(n), on obtient :

$$\begin{aligned}\hat{t}_{y\pi} &= \frac{N}{n}\sum_{k\in S}y_k,\\ V_p\left[\hat{t}_{y\pi}\right] &= N^2\,\frac{1-f}{n}S_y^2,\\ CV_p\left[\hat{t}_{y\pi}\right] &\simeq \frac{1}{\sqrt{n}}\,\frac{S_y}{\mu_y}.\end{aligned}$$

Calcul de précision

La formule précédente montre que pour un SRS, la variance du π-estimateur est en $O(n^{-1})$. On retiendra que pour un plan de sondage quelconque, la variance est du même ordre de grandeur.

Comme le π-estimateur est sans biais, on a également

$$EQM_p\left[\hat{t}_{y\pi}\right] = O(n^{-1}),$$

d'où l'ordre de grandeur en probabilité de $\hat{t}_{y\pi}$:

$$\hat{t}_{y\pi} \quad = \quad O_p\left(n^{-1/2}\right).$$

Redressement d'un estimateur

Principe

Dans l'estimateur direct

$$\hat{t}_{y\pi} = \sum_{k\in S} \frac{y_k}{\pi_k} = \sum_{k\in S} d_k y_k,$$

les **poids de sondage** d_k dépendent de l'information auxiliaire mobilisée au moment de l'échantillonnage :

$$\begin{array}{ll} \text{SRS} & \Rightarrow \quad d_k = N/n, \\ \text{SRS stratifié} & \Rightarrow \quad d_k = N_h/n_h \text{ pour } k \in U_h. \end{array}$$

Il se peut qu'une partie de l'information auxiliaire n'ait pas été utilisée au moment de la sélection de l'échantillon, ou qu'elle n'ait pas été disponible. On peut le faire au stade de l'estimation, en **redressant** le π-estimateur

$$\text{Poids de sondage } d_k \Rightarrow \text{Poids redressés } w_k$$

Principe

On dit que l'on **redresse** l'échantillon lorsque l'on modifie le système de pondérations associé à S afin de respecter une **information auxiliaire**.

On parle d'information auxiliaire lorsque l'on dispose d'une information connue **sur l'ensemble de la population**.

Exemples :

- Chiffre d'affaire total des entreprises d'un secteur d'activité,
- Répartition par sexe et par âge d'une population d'individus.

Estimateur par le ratio

On suppose connu le total t_x d'une variable auxiliaire (positive) x_k.
L'estimateur par le ratio est défini par

$$\hat{t}_{yR} = \hat{t}_{y\pi} \times \frac{t_x}{\hat{t}_{x\pi}} = \sum_{k \in S} w_k y_k$$

avec $w_k = d_k \times \frac{t_x}{\hat{t}_{x\pi}}$.

Principe : si l'estimateur direct surestime (ou sous-estime) le total t_x d'un facteur α, on suppose que le même facteur de surestimation (ou de sous-estimation) s'applique à l'estimation du total t_y.

L'estimation par le ratio sera efficace si la variable d'intérêt est approximativement proportionnelle à x.
Exemple : enquête auprès d'entreprises, avec redressement sur la variable d'effectif salarié.

Propriétés de l'estimateur par le ratio

On peut le réécrire sous la forme

$$\begin{aligned}\hat{t}_{yR} &= \hat{t}_{y\pi} + \hat{R}_\pi \, (t_x - \hat{t}_{x\pi}) \\ &= \hat{t}_{y\pi} + R \, (t_x - \hat{t}_{x\pi}) + (\hat{R}_\pi - R) \, (t_x - \hat{t}_{x\pi}),\end{aligned}$$

avec $\hat{R}_\pi = \hat{t}_{y\pi}/\hat{t}_{x\pi}$. Le biais de l'estimateur par le ratio est en $O(n^{-1})$, donc négligeable pour de grandes tailles d'échantillons.

Sa variance est approximativement donnée par

$$\begin{aligned}V_p\left[\hat{t}_{yR}\right] &\simeq V_p\left[\hat{t}_{y\pi} + R \, (t_x - \hat{t}_{x\pi})\right] \\ &= V_p\left[\hat{t}_{E\pi}\right]\end{aligned}$$

avec $E_k = y_k - R \, x_k$ et $R = t_y/t_x$. La variance est en $O(n^{-1})$. Elle sera réduite par rapport à l'estimateur direct si les variables y et x sont approximativement proportionnelles.

Estimateur par le ratio : cas du SRS

Pour un SRS, on obtient

$$\hat{t}_{yR} = t_x \times \frac{\bar{y}}{\bar{x}}.$$

Son biais est approximativement donné par

$$B_p\left[\hat{t}_{yR}\right] \simeq -N\frac{1-f}{n}\frac{S_{xy} - R\,S_x^2}{\mu_x}.$$

La variance est approximativement donnée par

$$\begin{aligned} V_p\left[\hat{t}_{yR}\right] &\simeq N^2\,\frac{1-f}{n}\,S_E^2 \\ &= N^2\,\frac{1-f}{n}\,\left[S_y^2 - 2RS_{xy} + R^2S_x^2\right]. \end{aligned}$$

Estimateur par la régression

Pour le vecteur $\mathbf{x}_k = [x_{1k}, \ldots, x_{pk}]^T$ de variables auxiliaires, les totaux $t_{\mathbf{x}} = [t_{x1}, \ldots, t_{xp}]^T$ sur U sont supposés connus.

L'utilisation de l'information auxiliaire apportée par le vecteur $\mathbf{x}_k$ est motivée par le modèle

$$y_k = \beta^T \mathbf{x}_k + \epsilon_k \quad \text{avec} \quad \begin{array}{c} E_m[\epsilon_k] = 0, \\ Cov_m[\epsilon_k, \epsilon_l] = \sigma_k^2 \, \delta_{kl}. \end{array} \tag{1}$$

L'**estimateur par la régression généralisée** (GREG) est donné par

$$\hat{t}_{y,greg} = \hat{\mathbf{b}}_\pi^T \, t_{\mathbf{x}} + \sum_{k \in S} d_k \, e_k$$

avec $\hat{\mathbf{b}}_\pi = \left[\sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^T}{\sigma_k^2 \, \pi_k} \right]^{-1} \sum_{k \in S} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \, \pi_k}$ et $e_k = y_k - \hat{\mathbf{b}}_\pi^T \, \mathbf{x}_k$.

Estimateur par la régression

L'estimateur GREG se compose donc de deux termes :

- $\hat{\mathbf{b}}_\pi^T \, t_\mathbf{x}$ est une prédiction du total t_y,
- $\sum_{k \in S} d_k \, e_k$ estime l'erreur totale de prédiction.

On peut encore le réécrire sous la forme :

$$\begin{aligned} \hat{t}_{y,greg} &= \hat{t}_{y\pi} + \hat{\mathbf{b}}_\pi^T \left[t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi} \right] \\ &= \sum_{k \in S} w_k y_k \end{aligned}$$

avec $w_k = d_k \left\{ 1 + \left[t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi} \right]^T \left[\sum_{k \in S} d_k \mathbf{x}_k \mathbf{x}_k^T \right]^{-1} \mathbf{x}_k \right\}$ les poids redressés .

Le coefficient $\hat{\mathbf{b}}_\pi$ joue un rôle de curseur : l'estimateur de base est d'autant plus modifié que le lien entre y et $\mathbf{x}$ est fort.

Propriétés de l'estimateur par la régression

On peut le réécrire sous la forme

$$\begin{aligned}\hat{t}_{y,greg} &= \hat{t}_{y\pi} + \hat{\mathbf{b}}_\pi^T \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi}\right] \\ &= \hat{t}_{y\pi} + \mathbf{b}^T \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi}\right] + \left[\hat{\mathbf{b}}_\pi - \mathbf{b}\right]^T \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi}\right]. \quad (2)\end{aligned}$$

Le biais de l'estimateur par la régression est en $O(n^{-1})$, donc négligeable pour de grandes tailles d'échantillons si le nombre de variables auxiliaires reste faible.

Sa variance est approximativement donnée par

$$\begin{aligned}V_p\left[\hat{t}_{y,greg}\right] &\simeq V_p\left[\hat{t}_{y\pi} + \mathbf{b}^T \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi}\right]\right] \\ &= V_p\left[\hat{t}_{E\pi}\right]\end{aligned}$$

avec $E_k = y_k - \mathbf{b}^T\mathbf{x}_k$. La variance est en $O(n^{-1})$, et sera réduite si les covariables $\mathbf{x}_k$ sont bien explicatives.

Estimateur par la régression : cas du SRS

On suppose que l'on utilise les variables auxiliaires $\mathbf{x}_k = [1, x_k]^T$, de totaux connus.

On obtient

$$\hat{b}_2 = \frac{\sum_{k\in S}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k\in S}(x_k - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \qquad \hat{b}_1 = \bar{y} - \hat{b}_2\ \bar{x}$$

et

$$\hat{t}_{y,greg} \quad = \quad \hat{t}_{y\pi} + \frac{s_{xy}}{s_x^2}\left[t_x - \hat{t}_{x\pi}\right].$$

En notant $\rho = \frac{S_{xy}}{S_x S_y}$ le coefficient de corrélation linéaire, on a :

$$V_p\left[\hat{t}_{y,greg}\right] \quad \simeq \quad N^2 \frac{1-f}{n} S_y^2 (1-\rho^2).$$

Application : estimateur post-stratifié

On suppose que la population U est partitionnée selon des groupes $U_1, \ldots, U_H$ dont les tailles $t_{\mathbf{x}} = [N_1, \ldots, N_H]^T$ sont connues après enquête.

Le modèle correspondant est : $y_k = \beta_h + \epsilon_k$ pour $k \in U_h$.

On obtient alors comme cas particulier de l'estimateur GREG l'estimateur post-stratifié

$$\hat{t}_{y,post} = \sum_{h=1}^{H} \hat{t}_{y_h} \frac{N_h}{\hat{N}_h},$$

avec $\hat{N}_h$ l'estimateur de la taille de la post-strate, et $\hat{t}_{y_h}$ l'estimateur du total sur la post-strate.

Estimateur par calage

On peut voir alternativement chacune des méthodes de redressement précédentes comme la recherche de nouveaux poids w_k qui

1. **restent proches** des poids de départ d_k,
2. **vérifient les équations de calage**

$$\sum_{k \in S} w_k \mathbf{x}_k = t_{\mathbf{x}}.$$

Plus formellement, on résoud le problème suivant :

$$\min_{w_k} \sum_{k \in s} d_k G\left(\frac{w_k}{d_k}\right) \text{ s.c. } \sum_{k \in s} w_k \mathbf{x}_k = t_{\mathbf{x}}$$

où G désigne une **fonction de distance** vérifiant certaines conditions de régularité.

Estimateur par calage

La résolution du problème d'optimisation conduit à :

$$w_k = d_k F[\lambda^T \mathbf{x}_k]$$

avec F la fonction inverse de G', et à l'estimateur calé

$$\hat{t}_{yw} = \sum_{k \in S} w_k y_k.$$

Les propriétés de l'estimateur par calage sont asymptotiquement les mêmes que celles de l'estimateur par la régression, d'où un biais en $O(n^{-1})$ et :

$$V_p\left[\hat{t}_{yw}\right] \simeq \sum_{k,l \in U} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \Delta_{kl}$$

où $E_k = y_k - \mathbf{b}^T \mathbf{x}_k$.

Résumé

Les estimateurs présentés précedemment (que l'on qualifiera d'estimateurs directs dans la suite) sont :

- sans biais dans le cas du π-estimateur,
- avec un très faible biais (en $O(n^{-1})$) dans le cas des estimateurs redressés.

Pour chacun d'eux, la variance est en $O(n^{-1})$; on obtiendra donc une estimation de bonne qualité si la taille d'échantillon est grande.

Dans le cas d'un estimateur redressé, la variance peut également être réduite de façon importante si on dispose de variables auxiliaires bien explicatives.

Estimation sur domaine

Estimation sur domaine

Lorsqu'un échantillon S est sélectionné aléatoirement dans une population U, il est fréquent qu'on l'utilise non seulement pour produire des estimations sur la population entière, mais également sur des sous-populations.

On parle alors d'*estimation sur domaine*. Il peut s'agir en particulier d'un domaine :

- géographique : individus vivant dans une région donnée,
- socio-démographique : individus de moins de 30 ans ; individus au chômage,
- temporel : individus présents sur le territoire français à une date donnée.

Estimation sur domaine

On note $U_d \subset U$ le domaine d'intérêt. On s'intéresse à l'estimation d'un total dans le domaine :

$$t_{yd} = \sum_{k \in U_d} y_k.$$

En remarquant que t_{yd} peut être vu comme le total sur U de la variable $z_k = y_k \ 1(k \in U_d)$, on obtient son estimateur de Horvitz-Thompson

$$\begin{aligned} \hat{t}_{yd} &= \sum_{k \in S} \frac{z_k}{\pi_k} \\ &= \sum_{k \in S_d} \frac{y_k}{\pi_k} \end{aligned}$$

avec $S_d = S \cap U_d$ la partie de l'échantillon S sélectionnée dans U_d. La taille de S_d (aléatoire) est notée n_d.

Interprétation

Seule la partie de l'échantillon qui tombe dans U_d est exploitée. Pour l'estimation de t_{yd}, on a donc "gaspillé" des ressources pour collecter une information "inutile" sur $U \setminus U_d$.

L'estimateur $\hat{t}_{yd}$ est un estimateur sans biais de t_{yd}. Pour un plan de taille fixe, sa variance est donnée par :

$$V_p\left[\hat{t}_{yd}\right] = -\frac{1}{2}\sum_{k\neq l\in U}\left(\frac{y_k\ 1(k\in U_d)}{\pi_k} - \frac{y_l\ 1(l\in U_d)}{\pi_l}\right)^2 \Delta_{kl}.$$

On a une source d'alea supplémentaire qui provient de la taille aléatoire d'échantillon n_d.

Estimation sur domaine : cas du SRS

Le π-estimateur de t_{yd} peut se réécrire sous la forme

$$\hat{t}_{yd} \quad = \quad \hat{N}_d \, \bar{y}_d,$$

avec $\bar{y}_d = \frac{1}{n_d} \sum_{k \in S_d} y_k$ la moyenne simple sur le sous-échantillon S_d et $\hat{N}_d = \frac{N \; n_d}{n}$ l'estimateur de la taille N_d du domaine U_d.

Si le taux de sondage est faible, la variance est approximativement donnée par

$$V_p \left[\hat{t}_{yd} \right] \quad \simeq \quad N_d^2 \frac{1}{E_p[n_d]} \left[S_{yd}^2 + \mu_{yd}^2 \left(1 - \frac{N_d}{N} \right) \right].$$

Estimation sur domaine : cas du SRS

Cette variance est en $O(E_p[n_d]^{-1})$, et dépend de la taille d'échantillon attendue dans le domaine. Si le domaine d'intérêt est petit, la taille (moyenne) d'échantillon sélectionnée dans ce domaine sera faible et la précision sera mauvaise.

Le tableau suivant (Ardilly, 2005) donne, pour différents domaines définis par la PCS, la taille globale d'échantillon nécessaire pour obtenir dans le domaine un CV de 5% :

PCS	Clergé	Cadres fonction publique	Professions scientifiques	Agriculteurs retraités	Ouvriers retraités
p_d	0.001	0.005	0.01	0.02	0.05
n	400 000	80 000	40 000	20 000	8 000

Estimation d'une moyenne

La moyenne $\mu_{yd} = \frac{t_{yd}}{N_d}$ sur le domaine peut être estimée selon un principe de substitution ("plug-in") par

$$\tilde{\mu}_{yd} = \frac{\hat{t}_{yd}}{\hat{N}_d},$$

avec $\hat{N}_d = \sum_{k \in S_d} \frac{1}{\pi_k}$ l'estimateur de la taille du domaine.

Le biais de $\tilde{\mu}_{yd}$ est en $O[E_p(n_d)^{-1}]$. Il est généralement négligeable, sauf pour des domaines très petits. Dans le cas d'un SRS, sa variance est approximativement donnée par :

$$V_p\left[\tilde{\mu}_{yd}\right] \simeq \left[\frac{1}{E_p[n_d]} - \frac{1}{N_d}\right] S_{yd}^2.$$

Utilisation d'information auxiliaire

Redressement par le ratio

Si la taille N_d du domaine est connue au moment de l'estimation, on peut améliorer l'estimateur $\hat{t}_{yd}$ en opérant un redressement par le ratio. On obtient :

$$\tilde{t}_{yd} = \hat{t}_{yd} \times \frac{N_d}{\hat{N}_d},$$

avec $\hat{N}_d = \sum_{k \in S_d} \frac{1}{\pi_k}$.

Dans le cas d'un SRS, on obtient

$$\tilde{t}_{yd} = N_d \, \bar{y}_d.$$

Pour étudier les propriétés de cet estimateur, on va raisonner conditionnellement à la taille d'échantillon n_d effectivement obtenue dans le domaine (approche conditionnelle).

Redressement par le ratio

On obtient

$$E_p\left[\tilde{t}_{yd}\,|n_d\right] = t_{yd} \quad \text{et} \quad V_p\left[\tilde{t}_{yd}\,|n_d\right] = N_d^2\left(\frac{1}{n_d} - \frac{1}{N_d}\right)S_{yd}^2$$

avec

$$S_{yd}^2 \quad = \quad \frac{1}{N_d - 1}\sum_{k\in U_d}\left(y_k - \mu_{yd}\right)^2.$$

L'estimateur $\tilde{t}_{yd}$ est non biaisé conditionnellement. Sa variance est en $O(n_d^{-1})$.

Si le domaine n'est pas sur-représenté lors de l'échantillonnage, la précision se détériore quand la taille du domaine diminue.

Redressement par le ratio

Intuitivement, l'approche conditionnelle colle mieux aux données (on raisonne par rapport à la taille d'échantillon effectivement observée dans le domaine). La variance conditionnelle est cependant difficile à calculer pour un plan de sondage quelconque. Notons également que l'estimateur direct $\hat{t}_{yd}$ est biaisé si l'on raisonne non conditionnellement.

L'estimateur par le ratio sera efficace si la variable d'intérêt y est peu dispersée dans le domaine.

Il est important de connaître au moment de l'échantillonnage les domaines les plus importants, afin de sélectionner dedans un échantillon de taille suffisante. Dans le cas contraire, la variance augmente de façon mécanique quand la taille du domaine diminue.

Redressement par la régression (1)

Si on connaît les totaux $t_{\mathbf{xd}}$ de variables auxiliaires $\mathbf{x}_k$ dans le domaine U_d, on peut caler l'estimateur direct $\hat{t}_{yd}$ sur ces totaux.

Avec l'estimateur par la régression, on obtient par exemple

$$\begin{aligned}\hat{t}_{y,greg1} &= \hat{t}_{yd} + \hat{b}_d^T \left[t_{\mathbf{xd}} - \hat{t}_{\mathbf{xd}\pi}\right] \\ &= \sum_{k\in S_d} w_{1k} y_k,\end{aligned}$$

avec $\hat{\mathbf{b}}_d = \left[\sum_{k\in S_d} \frac{\mathbf{x}_k \mathbf{x}_k^T}{\sigma_k^2\ \pi_k}\right]^{-1} \sum_{k\in S_d} \frac{\mathbf{x}_k y_k}{\sigma_k^2\ \pi_k}$ le vecteur des coefficients de régression estimé sur l'échantillon S_d.

Les poids w_{1k} vérifient les équations de calage

$$\sum_{k\in S_d} w_{1k}\mathbf{x}_k = t_{\mathbf{xd}}.$$

Redressement par la régression (1)

Le biais de l'estimateur $\hat{t}_{y,greg1}$ est en $O(n_d^{-1})$. Il est donc négligeable, sauf pour de très petits échantillons.

La variance de l'estimateur $\hat{t}_{y,greg1}$ est également en $O(n_d^{-1})$. L'ordre de grandeur de la variance peut donc être conséquent si le domaine est petit.

Cependant, la variance est donnée par la variable de résidus $E_k = y_k - \mathbf{b}_d^T \mathbf{x}_k$. Elle peut donc être très fortement diminuée si on dispose d'une information auxiliaire dans le domaine bien explicative.

Redressement par la régression (2)

Sans incorporer d'information auxiliaire spécifique au domaine, on peut également utiliser les poids d'estimation correspondant à un calage sur les totaux $t_{\mathbf{x}}$ dans la population U.

On obtient

$$\hat{t}_{y,greg2} = \sum_{k \in S_d} w_{2k} y_k,$$

où les poids w_{2k} vérifient les équations de calage dans la population

$$\sum_{k \in S} w_{2k} \mathbf{x}_k = t_{\mathbf{x}}.$$

C'est l'estimation que l'on obtient dans une enquête pour les domaines sur lesquels on n'effectue pas de redressement.

Redressement par la régression (2)

Intuitivement, ce dernier estimateur devrait être moins efficace que $\hat{t}_{y,greg1}$ qui utilisait une information spécifique au domaine. Ici, la variance est donnée par la variable de résidus

$$\begin{aligned} E_k &= y_k \, 1(k \in U_d) - \mathbf{b}^T \, \mathbf{x}_k \\ &= \begin{cases} y_k - \mathbf{b}^T \, \mathbf{x}_k & \text{si} \quad k \in U_d, \\ -\mathbf{b}^T \, \mathbf{x}_k & \text{si} \quad k \notin U_d. \end{cases} \end{aligned}$$

En particulier, ces résidus peuvent être largement négatifs.

Cet estimateur a cependant un avantage : son biais est en $O(n^{-1})$, et sera donc négligeable si l'échantillon S est grand, et même si le domaine est très petit.

En résumé

Dans un domaine, le biais d'un estimateur direct est généralement négligeable (sauf pour de tous petits échantillons). La variance est en $O(n_d^{-1})$, et se détériore pour les petits domaines.

La variance peut être réduite par redressement de l'estimateur direct, en incorporant de l'information auxiliaire bien explicative de la variable d'intérêt. Cependant, l'ordre de grandeur de la variance reste le même.

Petits domaines
Modélisation implicite

Estimateur direct modifié

Estimateur direct modifié

Une première idée consiste à partir de l'estimateur par la régression (1) utilisant de l'information auxiliaire dans le domaine. On remplace le coefficient de régression $\hat{\mathbf{b}}_d$ estimé sur S_d par le coefficient de régression $\hat{\mathbf{b}}$ estimé sur l'échantillon entier S.

On obtient :

$$\hat{t}_{yd,mod} = \hat{t}_{yd} + \hat{\mathbf{b}}^T \left[t_{\mathbf{xd}} - \hat{t}_{\mathbf{xd}\pi}\right].$$

Le biais est négligeable si la taille n d'échantillon est grande. La variance est en $O(n_d^{-1})$.

La variance est donnée par la variable de résidus $E_k = y_k - \mathbf{b}^T \mathbf{x}_k$.

Estimateur direct modifié

Cet estimateur repose sur l'hypothèse que le lien entre la variable y et les covariables $\mathbf{x}$ est le même sur le domaine U_d et dans la population entière, autrement dit :

$$\mathbf{b}_d \simeq \mathbf{b}.$$

Si c'est le cas, et si les variables auxiliaires sont bien explicatives, la variance sera réduite.

Néanmoins, elle sera généralement plus forte qu'avec les "vrais" résidus $y_k - \mathbf{b}_d^T \mathbf{x}_k$.

Estimation synthétique

Principe

On parle d'estimation synthétique lorsqu'un estimateur direct pour la population entière (ou pour un plus grand domaine) est considéré comme fiable, et que cet estimateur est utilisé pour produire une estimation sur de plus petits domaines.

Les estimateurs synthétiques reposent sur une modélisation implicite : une relation observée dans la population (ou dans un grand domaine) est supposée être la même dans le petit domaine pour lequel on veut produire une estimation.

Estimation synthétique sans information auxiliaire

Supposons que l'on s'intéresse à l'estimation de la moyenne μ_{yd} dans le domaine. En l'absence d'information auxiliaire, un estimateur possible est donné par :

$$\hat{\mu}_{yd,synt} = \frac{\hat{t}_{y\pi}}{\hat{N}}.$$

La variance de cet estimateur est en $O(n^{-1})$, et sera donc faible même si le domaine est petit. Son biais ne décroît pas avec la taille d'échantillon, et vaut approximativement

$$B_p\left[\hat{\mu}_{yd,synt}\right] \simeq \mu_y - \mu_{yd}.$$

Estimation synthétique sans information auxiliaire

Le biais s'annule donc si le comportement moyen dans le domaine est le même que dans la population entière.

Cette hypothèse peut être peu réaliste. Une autre possibilité consiste à ne pas utiliser l'échantillon complet pour produire l'estimateur synthétique, mais l'échantillon tiré dans un grand domaine intermédiaire entre le petit domaine d'intérêt et l'ensemble de la population.
Exemple : estimation du taux de chômage d'une commune, en partant de l'hypothèse que le taux de chômage est constant au sein de la zone d'emploi de la commune.

En pratique, on doit trouver un compromis entre l'utilisation d'un sur-domaine suffisamment grand (afin que la variance soit fortement diminuée), mais pour lequel l'hypothèse semble réaliste.

Estimation synthétique avec information auxiliaire

Supposons maintenant que des totaux $t_{\mathbf{x}d}$ sur le domaine sont connus. Alors l'estimateur synthétique du total t_y est donné par

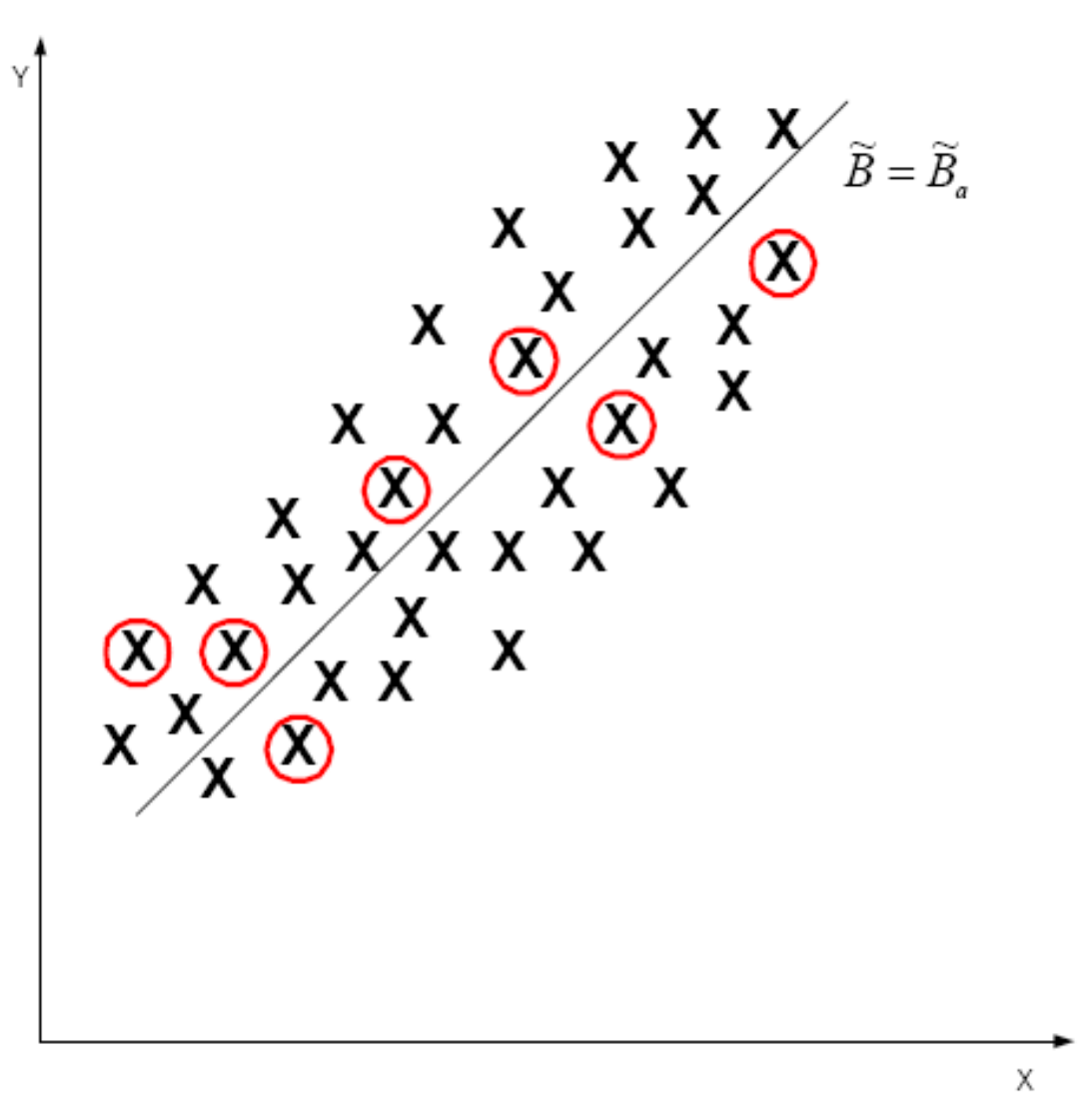
$$\hat{t}_{yd,synt} = \hat{\mathbf{b}}^T t_{\mathbf{x}d}.$$

La variance de cet estimateur est en $O(n^{-1})$. Son biais ne décroît pas quand la taille d'échantillon n augmente, et vaut approximativement

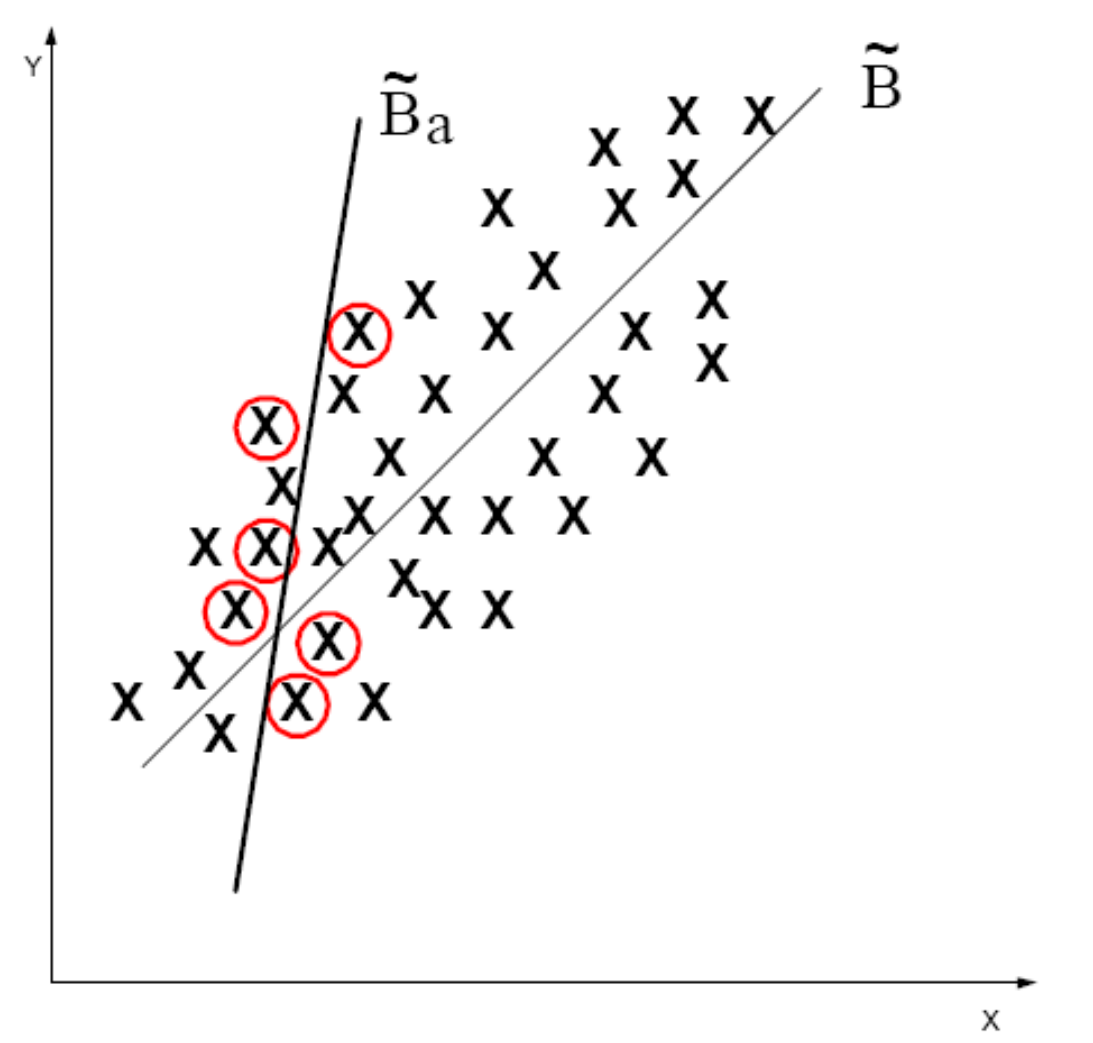
$$B_p\left[\hat{t}_{yd,synt}\right] \simeq \mathbf{b}^T t_{\mathbf{x}d} - t_{yd}.$$

Ce biais sera faible si les variables auxiliaires $\mathbf{x}_k$ sont bien explicatives de la variable y, **et** si la relation entre y et $\mathbf{x}$ est la même dans le domaine et dans le reste de la population ($\mathbf{b}_d \simeq \mathbf{b}$).

Exemple favorable (Ardilly, 2005)



Exemple défavorable (Ardilly, 2005)



Estimation synthétique : cas d'une post-stratification

Supposons que le domaine U_d soit partitionné selon un système de H post-strates, et que les effectifs $N_{dh},\ h = 1, \ldots, H$ dans le domaine et dans chaque post-strate soient connus.
Exemple : zone d'emploi dans laquelle les effectifs dans des classes d'âge sont connus.

Si on dispose d'estimateurs $\hat{t}_{yh}$ et $\hat{N}_h$ jugés fiables dans les post-strates complètes, on peut utiliser l'estimateur

$$\hat{t}_{y,dsynt} = \sum_{h=1}^{H} N_{dh} \frac{\hat{t}_{yh}}{\hat{N}_h}.$$

Exemple : estimation du nombre de chômeurs dans le domaine, en utilisant les effectifs estimés au niveau de la région.

Estimation synthétique : cas d'une post-stratification

Dans le cas où on s'intéresse à une proportion dans le domaine, y désigne une variable indicatrice. On connaît une estimation $\hat{p}_h = \frac{\hat{t}_{yh}}{\hat{N}_h}$ jugée fiable de la proportion dans la post-strate.

On peut utiliser l'estimateur synthétique de la proportion

$$\hat{p}_{dsynt} = \frac{\sum_{h=1}^{H} N_{dh}\hat{p}_h}{\sum_{h=1}^{H} N_{dh}}.$$

Application : enquête sur la santé

Enquête sur la natalité aux Etats-Unis, réalisée en 1980 sur un échantillon de $9\ 941$ naissances (Gonzalez et al., 1996). Estimation de la proportion de cas de jaunisse dans chaque état (vu comme un petit domaine), ici la Pennsylvanie.

La population est partitionnée en $H = 25$ post-strates, obtenue en croisant la race de la mère, sa classe d'âge et le rang de naissances.

Utilisation des proportions de cas de jaunisse estimés au niveau national dans chaque post-strate + nombres de naissances par état et par post-strate $N_{dg} \Rightarrow$ estimateur synthétique

$$\begin{aligned} \hat{p}_{dsynt} &= \frac{\sum_{h=1}^{H} N_{dh}\hat{p}_h}{\sum_{h=1}^{H} N_{dh}} = \frac{33\ 806}{156\ 799} \\ &= 21.6\% \ . \end{aligned}$$

Estimation de l'EQM

Dans le cas d'un estimateur sur petit domaine, on estime généralement l'EQM plutôt que la variance car le biais n'est pas nécessairement négligeable. D'autre part, la variance est souvent beaucoup plus faible que celle des estimateurs directs (on "importe de la force" depuis des domaines plus grands).

L'estimation de l'EQM pose problème : la variance peut s'estimer de façon habituelle si les propriétés du plan de sondage sont connues, mais le biais est souvent difficile à estimer.

Estimateur de l'EQM (1)

On peut utiliser l'identité

$$EQM_p\left[\hat{t}_{yd,s}\right] = E_p\left[\hat{t}_{yd,s}-\hat{t}_{yd}\right]^2 - V_p\left[\hat{t}_{yd,s}-\hat{t}_{yd}\right] + V_p\left[\hat{t}_{yd,s}\right],$$

en notant $\hat{t}_{yd,synt} \equiv \hat{t}_{yd,s}$, et où $\hat{t}_{yd}$ désigne un estimateur direct de t_{yd}.

On obtient l'estimateur

$$eqm_1\left[\hat{t}_{yd,s}\right] = \left[\hat{t}_{yd,s}-\hat{t}_{yd}\right]^2 - v\left[\hat{t}_{yd,s}-\hat{t}_{yd}\right] + v\left[\hat{t}_{yd,s}\right],$$

avec $v\left[\cdot\right]$ un estimateur de variance sans biais sous le plan de sondage.

Estimateurs de l'EQM (2)

L'estimateur de l'EQM proposé peut être très instable (et même prendre des valeurs négatives). Une solution possible consiste à moyenner l'EQM estimé sur une classe de m petits domaines. On obtient l'estimateur :

$$eqm_2\left[\hat{t}_{yd,s}\right] = \frac{1}{m}\sum_{i=1}^{m} eqm_1\left[\hat{t}_{yi,s}\right].$$

On peut également utiliser un estimateur de l'EQM simplifié, en remarquant que la variance de $\hat{t}_{yd,s}$ est généralement faible devant celle de l'estimateur direct $\hat{t}_{yd}$. On obtient :

$$eqm_3\left[\hat{t}_{yd,s}\right] = \left[\hat{t}_{yd,s} - \hat{t}_{yd}\right]^2 - v\left[\hat{t}_{yd}\right].$$

Estimation de l'EQM : cas du SRS

On sélectionne un échantillon S dans U selon un SRS. En l'absence d'information auxiliaire, l'estimateur direct de la moyenne μ_{yd} dans le domaine est $\hat{\mu}_{yd} = \frac{\hat{t}_{yd}}{\hat{N}_d}$. On utilise ici comme estimateur synthétique

$$\hat{\mu}_{yd,s} = \frac{\hat{t}_{y\pi}}{\hat{N}} = \bar{y}.$$

L'estimateur simplifié de l'EQM est donné par

$$eqm_3\left[\hat{t}_{yd,s}\right] = \left[\bar{y} - \hat{\mu}_{yd}\right]^2 - \left[\frac{1}{n_d} - \frac{1}{N_d}\right] s_{yd}^2,$$

avec $s_{yd}^2 = \frac{1}{n_d - 1} \sum_{k \in S_d} \left(y_k - \bar{y}_d\right)^2$.

Estimation composite

Principe

Une manière naturelle d'arbitrer entre le biais d'un estimateur synthétique $\hat{t}_{yd,s}$ et la variance d'un estimateur direct $\hat{t}_{yd}$ consiste à prendre une moyenne pondérée de ces deux estimateurs.

On obtient l'estimateur composite

$$\hat{t}_{yd,c} = \phi_d \, \hat{t}_{yd,s} + (1 - \phi_d) \, \hat{t}_{yd},$$

avec $\phi_d \in [0, 1]$.

Problème : comment choisir le coefficient ϕ_d?

Estimateur "optimal"

Si on suppose que la covariance entre $\hat{t}_{yd,s}$ et $\hat{t}_{yd}$ est faible, alors le choix optimal (au sens de la minimisation de l'EQM) est approximativement donné par

$$\phi_d = \frac{EQM_p\left[\hat{t}_{yd}\right]}{EQM_p\left[\hat{t}_{yd}\right] + EQM_p\left[\hat{t}_{yd,s}\right]}.$$

Dans l'estimation, on accorde donc d'autant plus de poids à un estimateur que celui-ci est précis.

En pratique, les EQM sont inconnues et doivent être estimées.

Estimateur "optimal"

En notant que

$$EQM_p\left[\hat{t}_{yd}\right] + EQM_p\left[\hat{t}_{yd,s}\right] \simeq E_p\left[\hat{t}_{yd,s} - \hat{t}_{yd}\right]^2,$$

on peut estimer ϕ_d par

$$\hat{\phi}_d = \frac{eqm\left[\hat{t}_{yd}\right]}{\left[\hat{t}_{yd,s} - \hat{t}_{yd}\right]^2}.$$

Cependant, cet estimateur est très instable. Une autre possibilité consiste à calculer ces coefficients $\hat{\phi}_d$ pour une classe de domaines jugés comparables, et à utiliser la valeur moyenne de ces estimateurs pour chaque domaine.

Estimateurs dépendant de la taille d'échantillon

Ces estimateurs composites utilisent un coefficient ϕ_d qui dépend de la taille N_d du domaine et de son estimateur $\hat{N}_d$. Grosso modo, on choisit d'utiliser l'estimateur direct si la taille effective d'échantillon dans le domaine n_d est jugée assez grande.

Plus précisément, on utilise par exemple le coefficient

$$\phi_d = \begin{cases} 1 & \text{si} \quad \frac{\hat{N}_d}{N_d} \geq \delta, \\ \frac{\hat{N}_d}{\delta \ N_d} & \text{si} \quad \frac{\hat{N}_d}{N_d} < \delta. \end{cases}$$

Rao (2003) suggère l'utilisation de l'estimateur GREG modifié comme estimateur direct, et de l'estimateur synthétique correspondant comme estimateur alternatif.

Estimation composite : cas du SRS

On sélectionne un échantillon S dans U selon un SRS. On a $\hat{N}_d = \frac{n_d}{n} N$. On peut utiliser comme estimateur composite

$$\begin{array}{lllll} \phi_d = 1 & \text{si} & f_d = \frac{n_d}{N_d} \geq \delta\ f & \Rightarrow & \hat{t}_y = \hat{t}_{yd} \\ \phi_d = \frac{f_d}{\delta\ f} & \text{si} & f_d < \delta\ f & \Rightarrow & \hat{t}_y = \frac{f_d}{\delta\ f}\ \hat{t}_{yd} + \left(1 - \frac{f_d}{\delta\ f}\right)\ \hat{t}_{yd,s} \end{array}$$

On utilise donc l'estimateur direct si la taille d'échantillon n_d est assez grande. Plus cette taille diminue, plus on donne de poids à l'estimateur synthétique.

Estimateur de James-Stein

Un autre choix d'estimateur composite consiste à utiliser, pour une classe de domaines fixée, le coefficient ϕ permettant d'obtenir la meilleure précision pour l'ensemble des domaines.

On obtient le coefficient (approximativement) optimal

$$\phi^* = \frac{\sum_{i=1}^m EQM_p\left[\hat{t}_{yi}\right]}{\sum_{i=1}^m \left[EQM_p\left(\hat{t}_{yi}\right) + EQM_p\left(\hat{t}_{yi,s}\right)\right]}.$$

On peut par exemple l'estimer par

$$\hat{\phi}^* = \frac{\sum_{i=1}^m eqm\left[\hat{t}_{yi}\right]}{\sum_{i=1}^m \left[\hat{t}_{yi} - \hat{t}_{yi,s}\right]^2}.$$

Application : mesure du chômage

Etude par simulation réalisée par Falorsi et al. (1994). Estimation de nombres de chômeurs dans $D = 14$ petits domaines de la région de Frioul. Ces domaines n'ont pas été pris en compte lors du plan de sondage.

Etude par simulations pour comparer les performances :

- d'un estimateur direct par post-stratification (POST) ;
- d'un estimateur synthétique de type post-stratifié (SYN) ;
- de l'estimateur composite optimal (COMP), où le coefficient ϕ_d est estimé à l'aide de données du Recensement ;
- de l'estimateur dépendant de la taille d'échantillon (SD), avec $\delta = 1$.

Application : mesure du chômage

Simulations de tirages de $B = 400$ échantillons, selon un tirage à deux degrés (UP=communes, US=ménages) avec stratification des unités primaires.

Indicateurs mesurés :

- Biais relatif moyen :

$$BRM = \frac{1}{14}\sum_{d=1}^{14}|RB_d| \quad \text{où} \quad RB_d = \frac{1}{400}\sum_{b=1}^{400}\frac{\hat{t}_{yd,b} - t_{yd}}{t_{yd}}.$$

- EQM relative moyenne :

$$EQMM = \frac{1}{14}\sum_{d=1}^{14}|RMSE_d| \quad \text{où} \quad RMSE_d = \frac{\sqrt{EQM_{d,sim}}}{t_{yd}}.$$

Résultats obtenus

Estimateur	BRM	EQMM
POST	1.75	42.08
SYN	8.97	23.80
COMP	6.00	23.57
SD	2.39	31.08

Commentaires :

- L'estimateur POST présente le plus faible biais, suivi par l'estimateur SD. L'estimateur SYN présente le biais le plus fort.
- Les estimateurs SYN et COMP présentent les EQM les plus faibles. A noter la bonne performance de SYN.

Petits domaines : modélisation explicite

Principe

Nous avons considéré pour l'instant que la seule source de variabilité de l'estimation était liée au sondage. L'apport d'information auxiliaire (par exemple, par redressement) permettait de réduire la variance des estimateurs.

Dans le cas d'une modélisation explicite, le paramètre d'intérêt θ_d lié au domaine est lui-même vu comme le résultat d'un processus aléatoire (alea dit de modèle). On peut réaliser une modélisation "agrégée" au niveau du domaine, ou une modélisation au niveau au niveau de l'individu.

Nous allons maintenant étudier des estimateurs issus d'une modélisation prenant en compte ces deux sources d'alea : on parle de modèle à effets aléatoire.

Modèle au niveau du domaine

Le modèle de Fay et Herriot (1979)

Soit $\theta_i = g(\mu_{yi})$ le paramètre à estimer sur le domaine U_i, $i = 1, \ldots, m$, où la fonction $g(\cdot)$ est connue. On suppose que chaque paramètre θ_i est issu du modèle

$$m : \theta_i = \mathbf{z}_i^T \beta + b_i v_i \text{ pour tout } i = 1, \ldots, m,$$

avec

- $\mathbf{z}_i$ un vecteur de covariables connu pour chaque domaine U_i ;
- β un vecteur (inconnu) ;
- b_i un réel (supposé connu) ;
- v_i une variable aléatoire centrée, de variance σ_v^2 (inconnue).

Les termes d'alea v_i sont supposés indépendants entre eux.

Interprétation

La valeur du paramètre θ_i est donnée en partie par des variables explicatives ($\mathbf{z}_i$), et en partie par un effet propre au domaine modélisé par le terme $b_i v_i$ (variabilité inter-petits domaines).

Les quantités intervenant dans ce modèle sont de nature différente : le vecteur $\mathbf{z}_i$ correspond à un effet fixe, et le terme v_i à un effet (supposé) aléatoire. La modélisation peut paraître discutable, mais elle a l'avantage de réduire la dimension du problème.

Notons que le paramètre θ_i n'est pas observé en pratique.

Alea d'échantillonnage

Sur la base des données d'enquête, on dispose d'autre part d'un estimateur direct $\hat{\theta}_i$ pour le paramètre θ_i. On a :

$$\hat{\theta}_i = \theta_i + e_i,$$

en notant e_i l'erreur d'échantillonnage, de variance ψ_i^2.

On suppose que les erreurs d'échantillonnage e_i sont sans biais sous le plan de sondage ; cette hypothèse peut être mise en défaut pour une fonction $g(\cdot)$ non linéaire, et pour une petite taille d'échantillon dans le domaine.

Modèle de Fay et Herriot

On suppose également que les termes d'erreur e_i sont indépendants entre eux. Cette hypothèse peut être peu réaliste, par exemple dans le cas d'un plan de sondage à plusieurs degrés, mais il est possible de l'affaiblir.

On suppose enfin que les termes d'erreur v_i et e_i sont indépendants deux à deux. Cette hypothèse est généralement réaliste, car le plan de sondage est souvent réalisé au niveau national sans tenir compte des spécificités locales.

On obtient finalement le modèle dit de Fay et Herriot (1979) :

$$\hat{\theta}_i \quad = \quad \mathbf{z}_i^T \ \beta + b_i v_i + e_i.$$

Modèle de Fay et Herriot

Rappelons que ce modèle empile deux aleas de nature différente:

- un alea de type modèle : $b_i \; v_i$,
- un alea de type sondage : e_i.

Le terme aléatoire global $b_i \; v_i + e_i$ est d'espérance nulle (sous le double mécanisme aléatoire), et de variance $b_i^2 \; \sigma_v^2 + \psi_i^2$.

Le modèle de Fay et Herriot peut se généraliser facilement à un paramètre multidimensionnel $\theta_i = (\theta_{1i}, \ldots, \theta_{Ki})$. On utilise alors le modèle :

$$\hat{\theta}_i \quad = \quad \mathbf{Z}_i^T \; \beta + V_i + E_i.$$

Extension : erreurs d'échantillonnage corrélées

Dans le cas du modèle multidimensionnel, les termes d'erreur V_i et E_i sont des vecteurs, et on note Σ_v^2 et Ψ leurs matrices de variance-covariance respectives.

Il peut être peu réaliste de supposer que la matrice Ψ est diagonale, i.e. que les estimateurs $\hat{\theta}_{li},\ l = 1, \ldots, K$ sont non corrélés.

On peut utiliser la matrice de variance-covariance associée au sondage, que l'on peut remplacer en pratique par une matrice estimée.

Extension : corrélations spatiales

On a supposé que les effets locaux v_i étaient indépendants. En pratique, cette hypothèse peut être peu crédible, par exemple si on s'intéresse à des domaines proches géographiquement.

On peut adapter la modélisation en postulant (par exemple) une covariance entre les aleas v_i et v_j de deux domaines de la forme

$$Cov(v_i, v_j) \quad = \quad \alpha \; e^{-\gamma \; d(i,j)},$$

en utilisant une mesure de distance $d(\cdot, \cdot)$ entre deux domaines.

On intégre donc un "effet de contagion", qui s'estompe avec la distance.

Extension : données répétées

Si on dispose de données sur plusieurs années (cas d'une enquête répétée dans le temps), on peut utiliser un modèle temporel de la forme

$$\begin{aligned} \hat{\theta}_{it} &= \theta_{it} + e_{it} \\ \theta_{it} &= \mathbf{z}_{it}^T \beta + b_i \, v_i + u_{it}, \end{aligned}$$

avec un effet aléatoire temporel u_{it}.

On peut par exemple utiliser une modélisation autorégressive pour l'effet temporel, de la forme

$$u_{it} = \rho \, u_{i,t-1} + \epsilon_{it}.$$

Modèle au niveau individuel

Principe

Soit y_{ik} la valeur de la variable d'intérêt pour l'individu k du domaine U_i. On postule pour la variable d'intérêt un modèle de la forme

$$m : y_{ik} = \mathbf{z}_{ik}^T \beta + v_i + e_{ik} \text{ pour tous } i = 1, \ldots, m \text{ et } k \in U_i,$$

avec

- $\mathbf{z}_{ik}$ un vecteur de covariables connu pour chaque individu k ;
- β un vecteur (inconnu) ;
- v_i une variable aléatoire centrée, de variance σ_v^2 (inconnue) ;
- e_{ik} un terme aléatoire centré, de variance $(K_{ik})^2 \sigma_e^2$.

Modèle au niveau individuel

Les termes v_i et e_{ik} sont supposés indépendants, et indépendants deux à deux.

Il est important de noter que le modèle est supposé vrai :

- pour tous les domaines considérés, et tous les individus de ces domaines,
- pour les individus de l'échantillon S sélectionné.

La dernière hypothèse suppose en particulier un mécanisme d'échantillonnag non informatif, i.e. que le mécanisme de sélection de l'échantillon soit indépendant des y_{ik}, conditionnellement aux valeurs $\mathbf{z}_{ik}$.

Modèle au niveau individuel

Comme pour le modèle au niveau domaine, on peut facilement étendre le modèle individuel au cas d'une variable d'intérêt multivariée.

En sommant la relation donnée par le modèle individuel m, on obtient d'autre part

$$\begin{aligned} \mu_{yi} &= \mathbf{z}_{ik}^T \beta + v_i + \bar{e}_i \\ &\simeq \mathbf{z}_{ik}^T \beta + v_i \end{aligned}$$

si le domaine est de taille N_i suffisamment grande. On retrouve donc une formulation voisine de celle obtenue avec une modélisation au niveau domaine.

Prise en compte d'un effet de grappe

L'hypothèse de résidus e_{ik} indépendants peut être peu réaliste, notamment dans le cas d'un plan de sondage à plusieurs degrés (risque de corrélation positive).

On peut modifier le modèle pour tenir compte d'un effet de grappe, de la façon suivante :

$$m : y_{ijk} = \mathbf{z}_{ijk}^T \beta + v_i + u_{ij} + e_{ijk}$$

pour un domaine U_i, une unité primaire j et un individu k. Le terme aléatoire u_{ij} permet d'incorporer un effet propre à l'unité primaire.

Variables qualitatives

Les modélisations précédentes ne sont pas adaptées au cas de variables qualitatives (en particulier, les hypothèses sur les résidus ne sont pas vérifiées). On peut introduire des effets aléatoires dans des modèles plus adaptés à l'étude de variables qualitatives.

On peut par exemple passer du modèle de régression logistique classique

$$\ln\left(\frac{p_{ik}}{1-p_{ik}}\right) = \mathbf{z}_{ik}^T \beta$$

au modèle logistique à effets aléatoires

$$\ln\left(\frac{p_{ik}}{1-p_{ik}}\right) = \mathbf{z}_{ik}^T \beta + v_i.$$

Estimation dans le modèle de Fay et Herriot

Principe

On ne revient pas ici sur les résultats théoriques concernant le calcul des estimateurs BLUP (Best Linear Unbiased Predictor) et EBLUP (Empirical BLUP) sous un modèle à effets aléatoires ; voir par exemple Rao (1995), chapitre 6.

Schématiquement, le BLUP fournit le meilleur prédicteur du paramètre d'intérêt (au sens de l'EQM), mais dépend de paramètres inconnus.

L'EBLUP s'obtient en remplaçant ces paramètres par des estimateurs.

Estimateur BLUP

Dans le cas du modèle de Fay et Herriot

$$\hat{\theta}_i = \mathbf{z}_i^T \beta + b_i v_i + e_i,$$

le BLUP est donné par

$$\tilde{\theta}_i^H = \mathbf{z}_i^T \tilde{\beta} + \gamma_i (\hat{\theta}_i - \mathbf{z}_i^T \tilde{\beta}),$$

avec

$$\gamma_i = \frac{b_i \sigma_v^2}{\psi_i + b_i^2 \sigma_v^2},$$

$$\tilde{\beta} = \left[\sum_{i=1}^m \frac{\mathbf{z}_i \mathbf{z}_i^T}{\psi_i + b_i^2 \sigma_v^2} \right]^{-1} \sum_{i=1}^m \frac{\mathbf{z}_i \hat{\theta}_i}{\psi_i + b_i^2 \sigma_v^2}.$$

Estimateur BLUP

Notons que cet estimateur n'est pas calculable en pratique, puisqu'il dépend de la variance σ_v^2 (inconnue).

Cet estimateur peut se réécrire sous une forme composite :

$$\tilde{\theta}_i^H = \gamma_i \underbrace{\hat{\theta}_i}_{\text{Est. direct}} + (1-\gamma_i) \underbrace{(\mathbf{z}_i^T \tilde{\beta})}_{\text{Est. synthétique}} .$$

Interprétation : si b_i ou σ_v^2 est petit, l'alea de modèle est faible et l'estimateur BLUP est presque égal à l'estimateur synthétique.

EQM de l'estimateur BLUP

L'estimateur BLUP est sans biais sous le double alea associé au modèle et à l'échantillonnage. Son biais sous le plan de sondage est approximativement donné par

$$B_p(\tilde{\theta}_i^H) \simeq [1-\gamma_i]\left[\mathbf{z}_i^T\ E_p(\tilde{\beta}) - \theta_i\right].$$

Son EQM est donnée par

$$\begin{aligned} EQM(\tilde{\theta}_i^H) &\simeq g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2), \text{ avec} \\ g_{1i}(\sigma_v^2) &= \frac{b_i^2\ \sigma_v^2\ \psi_i}{\psi_i + b_i^2\ \sigma_v^2}, \\ g_{2i}(\sigma_v^2) &= (1-\gamma_i)^2 \mathbf{z}_i^T \left[\sum_{i=1}^m \frac{\mathbf{z}_i\mathbf{z}_i^T}{\psi_i + b_i^2\ \sigma_v^2}\right]^{-1} \mathbf{z}_i. \end{aligned}$$

Le terme $g_{1i}(\sigma_v^2)$ est généralement prépondérant, si le nombre de domaines m est grand.

Estimateur EBLUP

Un estimateur de la variance σ_v^2 peut s'obtenir selon la méthode des moments, en utilisant l'identité

$$E\left[\sum_{i=1}^{m} \frac{(\hat{\theta}_i - \mathbf{z}_i^T \ \tilde{\beta})^2}{\psi_i + b_i^2 \ \sigma_v^2}\right] \quad = \quad m - p.$$

On obtient ensuite l'estimateur EBLUP en substituant dans le BLUP σ_v^2 par l'estimateur $\hat{\sigma}_{vm}^2$ ainsi obtenu.

Le calcul (et l'estimation) de l'EQM de cet estimateur est plus compliqué, et nécessite en particulier des hypothèses supplémentaires sur la normalité des résidus.

Bibliographie

- Ardilly, P. (2005). *Panorama des principales méthodes d'estimation sur petits domaines*. Actes des Journées de Méthodologie Statistique, Insee.
- Fay, R.E., and Herriot, R.A. (1979). *Estimation of income for small places: an application of James-Stein procedures to Census Data*. JASA, 74, p. 269-277.
- Falorsi, P.D., Falorsi, S., and Russo, A. (1994). *Empirical comparison of small area estimation methods for the Italian Labour Force Survey*. Survey Methodology, 20, p. 171-176.
- Gonzalez, M.E., Placek, P.J., and Scott, C (1996). Synthetic estimation of follow-back surveys at the National Center for Health Statistics. In Indirect Estimators in US Federal Programs, Springer-Verlag.
- Lahiri, P. (2011). *Bayesian Methods in Small Area Statistics*. ECAS, 2011.
- Rao, J.N.K (2003). *Small Area Estimation*. New-York, Wiley.