Approche inférentielle en théorie des sondages

Mohammed El Haj Tirari

Institut National de Statistique et d'Economie Appliquée

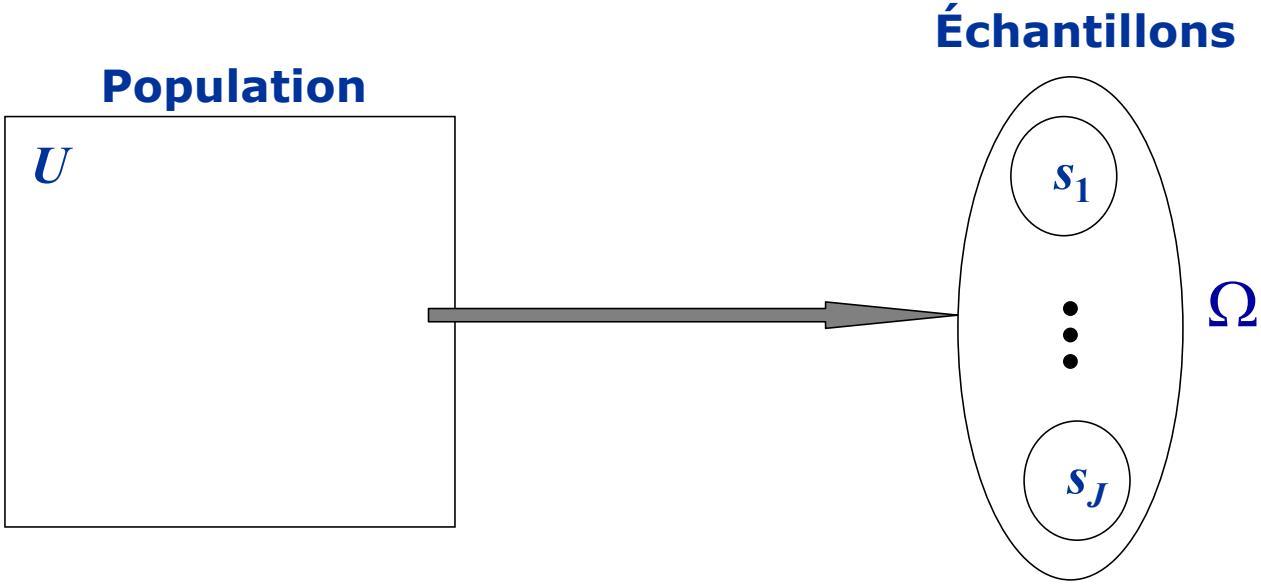
9^e Colloque Francophone sur les Sondages, Gatineau, Québec, Canada

Plan

Introduction

- Cette deuxième partie de l'atelier est consacrée à la présentation de l'approche inférentielle pour l'estimation en présence de non-réponse et pour l'estimation sur les petits domaines.
- A travers ces deux exemples, nous allons voir comment l'approche inférentielle permet de formaliser les hypothèses permettant de garantir les bonnes propriétés des estimateurs adoptés.
- Nous verons donc que l'information auxiliaire joue un rôle important dans l'élaboration de l'approche inférentielle :
  - $\hookrightarrow$ La validation des hypothèses de l'approche inférentielle nécessite l'utilisation de l'information auxiliaire de bonne qualité.

Approche basée sur la plan de sondage



Pour chaque échantillon s possible, on peut associer une probabilité de sélection $p(s)$

Approche basée sur la plan de sondage

- Cette approche permet de produire des estimateurs fiables pour des échantillons de grande taille mais ces résultats asymptotiques ne sont pas forcément adaptés pour des petits échantillons par exemple dans le cas des petits domaines.
- Elle ne permet pas de construire des estimateurs optimaux : il n'existe pas d'estimateur optimal (de variance minimum) quelque soit le plan de sondage et les valeurs de la variable d'intérêt Y.
- Sous cette approche, les calculs de variance sont compliqués pour les plans de sondage complexes.
- L'inférence de cette approche n'est plus adaptée quand le plan de sondage est perturbé, par exemple en présence de non-réponse.

Approche modèle

- Soit $U = \{1, \ldots, N\}$ une population de taille N à partir de laquelle on sélectionne un échantillon s de taille n.
- On s'intéresse à une variable d'intérêt $\mathbf{y} = (y_1, \ldots, y_N)'$ en ayant comme objectif l'estimation de son total :

$$t_y = \sum_{k \in U} y_k$$

- On suppose qu'on dispose de p variables auxilaires $X_1, \ldots, X_p$ dont les totaux

$$t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$$

sont connus avec, pour tout $k \in U$, $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kp}\right)'$.

Approche modèle

- La population est considérée comme "prélevée" à partir d'une superpopulation

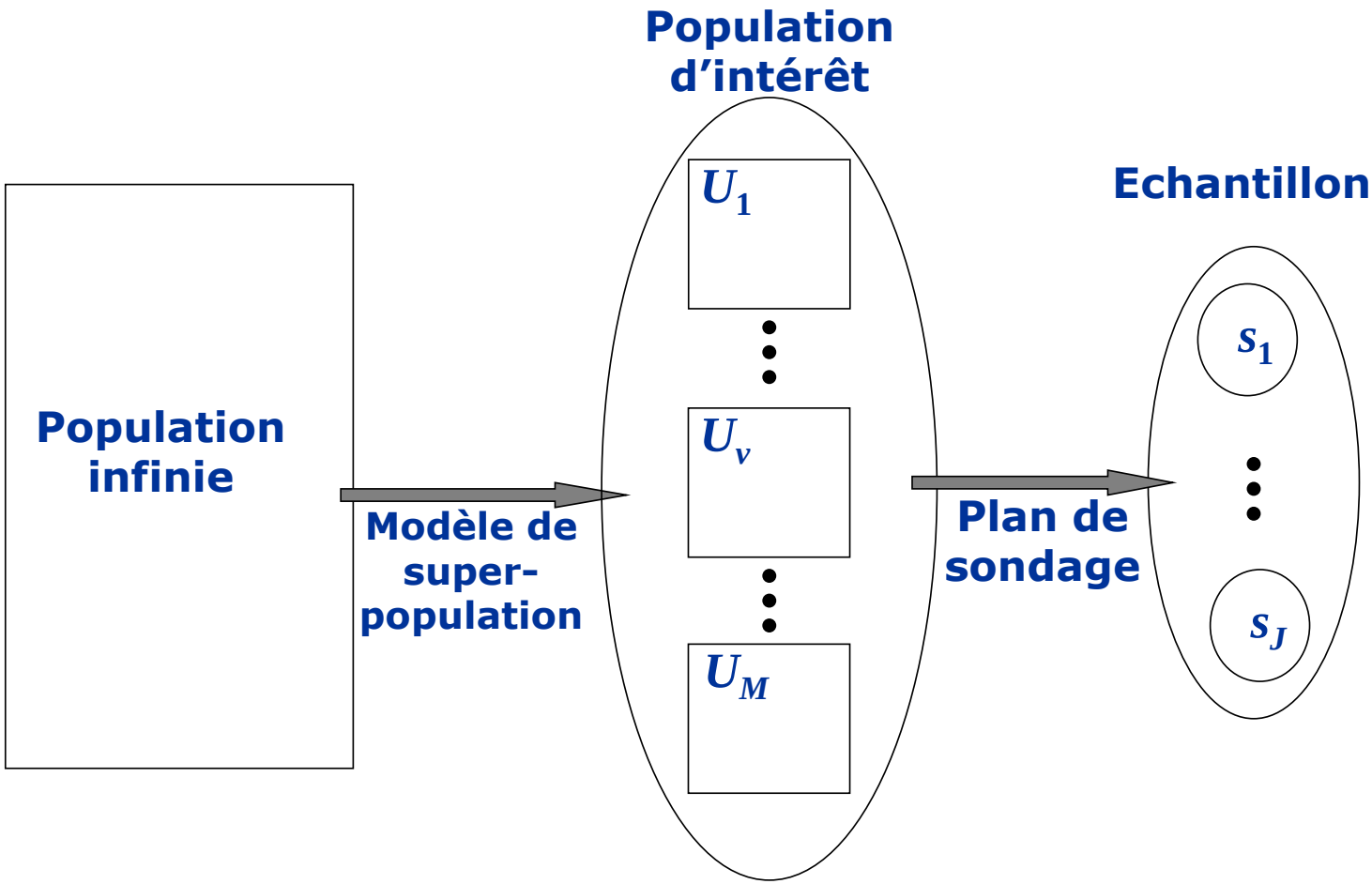
- Les valeurs de la variable d'intérêt sont les réalisations d'un vecteur aléatoire

$$(Y_1, \ldots, Y_k, \ldots, Y_N)'$$

dont la distribution est définie par un modèle stochastique appelé modèle de superpopulation

$\hookrightarrow$ l'échantillon est issu d'une double expérience aléatoire : une réalisation du modèle qui fournit la population et ensuite le tirage de l'échantillon

Approche modèle



Approche modèle

- Cette approche permet de produire des estimateurs optimaux selon l'EQM sous le modèle par exemple l'estimateur BLUE (Best Linear Unbiased Estimator).
- Elle permet d'élaborer une bonne inférence même dans le cas des petits échantillons (petits domaines).
- Elle permet d'inclure une modélisation pour la non-réponse.
- Cependant, elle peut produire des estimateurs imprécis dans le cas où le modèle est mal spécifie.
  - $\hookrightarrow$ C'est pour cela qu'il faut tenir compte aussi du plan de sondage (Approche basée sur la plan et sur le modèle).

Modèle de superpoulation

Sous l'approche basée sur le modèle, on suppose que les valeurs de $\mathbf{y}$ sont les réalisations du modèle de superpopulation ξ. Par exemple, ce modèle peut être :

$$y_k = \mathbf{x}_k\boldsymbol{\beta} + \epsilon_k$$

avec

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)',$$

$$E_\xi(\epsilon_k) = 0, \quad var_\xi(\epsilon_k) = \sigma^2 v_k^2 \text{ et } cov_\xi(\epsilon_k, \epsilon_l) = 0.$$

Les v_k^2 sont supposé connus avec

$$\sum_{k\in U} v_k = N$$

Pourquoi l'approche modèle ?

Les raisons derrière l'utilisation de l'approche modèle peuvent être résumées par :

1. Sous l'approche basée sur le plan de sondage, on ne dispose pas d'estimateur optimal de point de vue précision et efficacité.
2. L'impossibilité de trouver des solutions convenables sous l'approche basée sur le plan pour certaines problèmatiques : estimation sur petits domaines, traitement de la non-réponse, ...
3. La nécessité de mettre à profit les informations auxiliaires disponibles sur les unités de la population.

Biais et précision d'un estimateur sous l'approche modèle

Soit θ une fonction des valeurs de la variable d'intérêt Y et $\widehat{\theta}$ un estimateur de θ se basant sur l'échantillon sélectionné S.

1. Le biais de $\widehat{\theta}$ peut être donné par

$$Biais_{\xi}(\widehat{\theta}) = E_{\xi}(\widehat{\theta}) - \theta$$

2. La Précision de $\widehat{\theta}$ peut être mesurer par :

$$\begin{aligned} EQM_{\xi}\left(\widehat{\theta}\right) &= E_{\xi}\left(\widehat{\theta} - \theta\right)^2 \\ &= Var_{\xi}\left(\widehat{\theta}\right) + \left[Biais_{\xi}(\widehat{\theta})\right]^2 \end{aligned}$$

Biais et précision d'un estimateur sous l'approche modèle

On peut mesurer le biais et la précision en tenant compte également du plan de sondage et du modèle :

1. Le biais de $\widehat{\theta}$ sous le plan et le modèle peut être donné par

$$Biais_{p\xi}(\widehat{\theta}) = E_{\xi}\left[E_p(\widehat{\theta}) - \theta\right]$$

2. La Précision de $\widehat{\theta}$ peut être définie par :

$$\begin{aligned} EQM_{p\xi}\left(\widehat{\theta}\right) &= E_{\xi}\left[E_p\left(\widehat{\theta}-\theta\right)^2\right] \\ &= Var_{p\xi}\left(\widehat{\theta}\right) + \left[Biais_{p\xi}(\widehat{\theta})\right]^2 \end{aligned}$$

Exemples d'utilisation de l'approche basée sur le modèle

- Notons que l'utilisation de l'approche modèle a permis d'apporter des améliorations et des solutions à des problématiques rencontrées en sondage.
- Afin illustrer l'apport de l'approche modèle à l'élaboration des techniques d'estimation, nous considérons les deux domaines de recherches suivants :
  1. Le traitement de la non-réponse dans les enquêtes.
  2. L'estimation sur petits domaines.

Plan

Introduction

- Lors de la réalisation d'une enquête, il faut s'attendre à rencontrer inévitablement un certain taux de non-réponse.
- La non-réponse est un problème important dans les enquêtes car elle mène en général à des estimateurs ponctuels biaisés.
- Elle conduit à une précision plus faible puisque la taille de l'échantillon des répondants est plus petite que celle prévue au moment du tirage.
- Il existe deux types de non-réponse :
  - non-réponse totale : absence complète d'information sur une unité
  - non-réponse partielle : absence d'information sur certaines variables recueillies
- En général, les statisticiens d'enquête procèdent à un ajustement de poids (repondération) pour compenser la non-réponse totale et l'imputation pour corriger à la non-réponse partielle

Inférence en présence de la non-réponse

- L'inférence en présence de la non-réponse est déduite à partir de celle de l'échantillonnage à deux phases : la phase 1 est celle de la sélection de l'échantillon alors que l'échantillon des répondants peut être vu comme l'échantillon de deuxième phase.

Deux phases : $U \overset{phase1}{\rightarrow} s_1 \overset{phase2}{\rightarrow} s_2$

Non-réponse : $U \overset{phase1}{\rightarrow} s_1 \overset{phase2}{\rightarrow} s_r$

- On fait souvent appel à des hypothèses à propos du mécanisme de réponse en supposant un modèle de comportement des individus face à la non-réponse, c'est-à-dire des hypothèses sur la distribution des non-répondants.
- La validité de l'inférence qui en découle dépend totalement de la validité de modèle choisi.

Types de mécanismes de réponse

Supposons que les unités répondent indépendamment les unes des autres avec une probabilité p_k, c'est-à-dire que le mécanisme de réponse est décrit par :

$$a_k \sim B(1, p_k) \quad \text{pour tout} \quad k \in U$$

Pour modéliser la phase de la sélection de l'échantillon des répondants s_r à partir de s, on considère en général trois types de mécanismes de réponse :

1. uniforme (Missing Completely at Random - MCAR)
2. ignorable (Missing at Random - MAR)
3. non-ignorable (Not Missing at Random - NMAR)

Biais et variance en présence de la non-réponse

- Supposons que le paramètre à estimer est le total t_y d'une variable d'intérêt Y pour laquelle on a de la non-réponse qui va être traitée à l'aide d'une technique disponible (repondération ou imputation). L'erreur totale de l'estimateur $\widehat{t}_{y_{NR}}$ de t_y après traitement de la non-réponse est donnée par

$$\underbrace{\widehat{t}_{y_{NR}} - t_y}_{\text{erreur totale}} = \underbrace{\widehat{t}_y - t_y}_{\text{erreur due à l'échantillonnage}} + \underbrace{\widehat{t}_{y_{NR}} - \widehat{t}_y}_{\text{erreur due à la non-réponse}}$$

- L'étude du biais et de la variance impliquent donc deux mécanismes aléatoires :
  1. Le mécanisme d'échantillonnage
  2. Le mécanisme de non-réponse

Biais et variance en présence de la non-réponse

- Si on suppose que $\widehat{t}_y$ est un estimateur sans biais de t_y, le biais de $\widehat{t}_{y_{NR}}$ est donné par

$$Biais\left(\widehat{t}_{y_{NR}}\right) = E\left(\widehat{t}_{y_{NR}} - t_y\right) = E_p E_r\left(\widehat{t}_{y_{NR}} - \widehat{t}_y \mid s\right) = E_p\left(B_{NR}\right)$$

où B_{NR} est le biais conditionnel dû à la non-réponse et $E_r(\cdot)$ dénote l'espérance par rapport au mécanisme de réponse.

- En pratique, il est impossible de savoir si B_{NR} est presque nul car le mécanisme de réponse est inconnu.

- Quand $B_{NR} = 0$, la variance de $\widehat{t}_{y_{NR}}$ est donnée par

$$\begin{aligned} V\left(\widehat{t}_{y_{NR}}\right) &= V\left(\widehat{t}_{y_{NR}} - t_y\right) \\ &= \underbrace{V_p\left(\widehat{t}_y\right)}_{\text{variance due à l'échantillonnage}} + \underbrace{E_p V_r\left(\widehat{t}_{y_{NR}} - \widehat{t}_y \mid s\right)}_{\text{variance due à la non-réponse}} \end{aligned}$$

Biais et variance en présence de la non-réponse

- La mesure du biais et de la variance de $\widehat{t}_{y_{NR}}$ ne peut se faire qu'après avoir déterminer le mécanisme de non-réponse.
- L'étude du mécanisme de non-réponse requiert généralement une utilisation judicieuse de l'information auxiliaire disponible.
- La mise à profit de cette information auxiliaire nécessite la supposition d'un modèle pour expliquer le comportement des répondants afin de réduire le biais dû à la non-réponse.
- En présence de la non-réponse, l'utilisation des modèles est inévitable et la qualité des estimations est conditionnée par la disponibilité d'information auxiliaire de qualité.

Approches inférentielles en présence de la non-réponse

- Pour pouvoir étudier les propriétés des estimateurs utilisés en présence de la non-réponse, plusieurs approches inférentielles ont été proposées.
- Avant de corriger la non-réponse, on commence souvent par construire des classes de réponses homogènes composées d'unités ayant les mêmes caractéristiques et comportements vis-à-vis de la réponse.
  - ↪ Les hypothèses de l'inférence statistique se font à l'intérieur de ces classes.
- Dans ce qui suit, nous considérons le cas d'une seule classe et ceci par souci de simplicité.

Approches inférentielles en présence de la non-réponse

- Ainsi, les approches inférentielles les plus utilisées en pratique pour étudier les propriétés des estimateurs tenant compte de la non-réponse sont :
  - ∗ Approche basée sur le plan (BP) : le mécanisme de réponse est supposé uniforme à l'intérieur de chaque classe.
  - ∗ Approche basée sur le modèle (BM) : le mécanisme de réponse est supposé ignorale à l'intérieur de chaque classe et que l'inférence est fondée sur un modèle dont la forme générale est donnée par :

$$y_k = \mathbf{z}_k^{'}\boldsymbol{\beta} + \varepsilon_k$$

$$E(\varepsilon_k) = 0, \qquad E(\varepsilon_k^2) = \sigma_k^2 \quad et \quad E(\varepsilon_k \varepsilon_l) = 0$$

où $\mathbf{z}_k = (z_{1k}, \ldots, z_{pk})^{'}$ est le vecteur des valeurs des variables auxiliaires pour l'unité $k \in s$.

Approche inférentielle en présence de données imputées

Méthodes d'imputation

- On s'intéresse à l'estimation du total $t_y = \sum_{k \in U} y_k$ en se basant sur les données observées sur un échantillon aléatoire s, de taille n, sélectionné selon un plan de sondage $p(s)$.
- En présence de la non-réponse, il est donc impossible de calculer l'estimateur

$$\widehat{t}_y = \sum_{k \in s} w_k y_k \qquad \text{où} \qquad w_k = \frac{1}{\pi_k}$$

- Soient s_r, de taille r et s_{nr} , de taille m les deux ensembles des répondants et des non-répondants respectivement à l'item Y.

$$\hookrightarrow \qquad s = s_r \cup s_{nr} \qquad \text{et} \qquad r + m = n.$$

Méthodes d'imputation

- L'estimateur imputé de t_y est défini par

$$\widehat{t}_{yI} = \sum_{k \in s_r} w_k y_k + \sum_{k \in s_{nr}} w_k y_k^*$$

où y_k^* est la valeur imputée utilisée pour remplacer la valeur manquante y_k.

$\hookrightarrow$ la valeur de y_k^* dépend de la méthode d'imputation adoptée.

- Plusieurs méthodes d'imputation ont été proposées et le choix de celle qui est la plus convenable dépend principalement de l'information auxiliaire disponible et de son lien avec la variable d'intérêt pour laquelle on a de la non-réponse.
- Cette information auxiliaire peut être représentée par la connaissance pour les unités de l'échantillon s des valeurs de p variables $Z_1, \ldots, Z_p$.

Méthodes d'imputation

- On note que la majorité des méthodes d'imputation peut être représentée par le modèle suivant :

$$y_k = f(\mathbf{z}_k) + \varepsilon_k$$

$$E(\varepsilon_k) = 0, \qquad E(\varepsilon_k^2) = \sigma_k^2 \quad et \quad E(\varepsilon_k \varepsilon_l) = 0$$

où pour l'unité $k \in s$, $\mathbf{z}_k = (z_{1k}, \ldots, z_{pk})'$ est le vecteur des valeurs des variables auxiliaires disponibles.

$\hookrightarrow$ Méthodes déterministes : $y_k^* = \widehat{f}(\mathbf{z}_k)$.
Méthodes aléatoires : $y_k^* = \widehat{f}(\mathbf{z}_k) + \widehat{\varepsilon}_k$.

Méthodes d'imputation

- Exemples :
  - Imputation par la moyenne des répondants ($\mathbf{z}_k = z_k = 1$ et $\sigma_k^2 = \sigma^2$) :

$$y_k^* = \widehat{f}(\mathbf{z}_k) = \widehat{\beta}_r = \overline{y}_r$$

.

  - Imputation par le Ratio ($\mathbf{z}_k = z_k$ et $\sigma_k^2 = \sigma^2 z_k$) :

$$y_k^* = \widehat{f}(\mathbf{z}_k) = \widehat{\beta}_r z_k \quad \text{avec} \quad \widehat{\beta}_r = \frac{\overline{y}_r}{\overline{x}_r}$$

  - Imputation par la régression :

$$y_k^* = \widehat{f}(\mathbf{z}_k) = \mathbf{z}_k^{'}\widehat{\boldsymbol{\beta}}_r \quad \text{avec} \quad \widehat{\boldsymbol{\beta}}_r = \sum_{k \in s_r} \frac{w_k}{\sigma_k^2}\mathbf{z}_k\mathbf{z}_k^{'} \sum_{k \in s_r} \frac{w_k}{\sigma_k^2}\mathbf{z}_k y_k$$

L'approche inférentielle et l'estimateur imputé

- Les propriétés des estimateurs utilisés en l'absence de non-réponse ne sont plus valides en présence de non-réponse :
  - $\hookrightarrow$ Par exemple, l'estimateur d'Horvitz-Thompson $\widehat{t}_{y\pi}$ du total de la population t_y n'est plus sans biais quand on a de la non-réponse.
- Cependant, on peut montrer que le biais de l'estimateur imputé dépend de la validité des hypothèses de l'approche inférentielle adoptée et ceci à propos du :
  - $*$ mécanisme de réponse dans le cas de l'approche basée sur le plan ;
  - $*$ modèle dans le cas de l'approche basée sur le modèle.
- Ainsi, l'approche inférentielle n'est qu'une formalisation des hypothèses permettant de garantir les propriétés de l'estimateur imputé.

L'approche inférentielle et l'estimateur imputé

Pour illstrer le lien entre les hypothèses de l'approche inférentielle et les propriétés de l'estimateur imputé

$$\widehat{t}_{yI} = \sum_{k \in s_r} w_k y_k + \sum_{k \in s_{nr}} w_k y_k^*$$

du total de la population t_y où $w_k = \frac{1}{\pi_k}$, nous considérons les deux exemples suivants (Haziza, 2002) :

1. L'approche basée sur le plan et l'imputation par la Moyenne
2. L'approche basée sur le modèle et l'imputation par le Ratio

L'approche inférentielle et l'estimateur imputé

L'approche basée sur le plan et l'imputation par la Moyenne :

Soit $p_k = P(k \in s_r)$ la probabilité de répondre à l'item Y pour une unité k. On peut montrer que le biais de l'estimateur $\widehat{t}_{yI}$ dans le cas de l'utilisation de l'imputation par la moyenne est donné par :

$$Biais(\widehat{t}_{yI}) = E_p E_r(\widehat{t}_{yI} \mid s) - t_y \approx \frac{1}{\overline{P}} \sum_{k \in U} (p_k - \overline{P})(y_k - \mu_y)$$

où $\overline{P}$ est la moyenne dans la population des probabilités des réponses p_k.

On note que le biais de $\widehat{t}_{yI}$ est égal à 0 si la covariance entre la probabilité de réponse et la variable d'intérêt est nulle.

- ↪ Cette condition est satisfaite lorsque le mécanisme de réponse est uniforme $(p_k = p)$, ce qui correspond à l'hypothèse de l'approche basée sur le plan.
- ↪ L'information auxiliaire doit être utilisée de telle sorte à construire des classes d'imputation composées d'unités homogènes vis-à-vis de la réponse.

L'approche inférentielle et l'estimateur imputé

L'approche basée sur le modèle et l'imputation par le Ratio :

- Soit z_k les valeurs de la variable auxiliaire utilisée pour élaborer l'estimateur $\widehat{t}_{yI}$ utilisant l'imputation par le ratio. Sous l'approche basée sur le modèle, on a :

$$y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$$

$$E(\varepsilon_k) = 0, \qquad E(\varepsilon_k^2) = \sigma_k^2 \quad \text{et} \quad E(\varepsilon_k \varepsilon_l) = 0$$

- On peut montrer que le biais de l'estimateur $\widehat{t}_{yI}$ dans le cas de l'utilisation de l'imputation par le ratio est donné par :

$$Biais(\widehat{t}_{yI}) = E_r E_p E_m(\widehat{t}_{yI} - t_y) \approx \beta_0 \left[\frac{\mu_z}{\mu_{zp}} - 1\right]$$

où

$$\mu_z = \frac{1}{N}\sum_{k\in U} z_k \qquad \text{et} \qquad \mu_{zp} = \frac{\sum_{k\in U} p_k z_k}{\sum_{k\in U} p_k}$$

L'approche inférentielle et l'estimateur imputé

L'approche basée sur le modèle et l'imputation par le Ratio (suite) :

On note que le biais de $\widehat{t}_{yI}$ est égal à 0 si l'une des deux conditions est satisfaite :

1. Utiliser le modèle dont la constante est nulle ce qui convient le mieux pour l'imputation par le ratio :

$$\beta_0 = 0 \quad \Leftrightarrow \quad \text{utilisation du modèle} \quad y_k = \beta_1 z_k + \varepsilon_k$$

2. La covariance entre la probabilité de réponse et la variable auxiliaire Z doit être nulle :

$$\mu_z = \mu_{zp} \quad \Leftrightarrow \quad \frac{1}{\overline{P}} \sum_{k \in U} (p_k - \overline{P})(z_k - \mu_z) = 0$$

$\hookrightarrow$ Ceci survient dans le cas d'un mécanisme de réponse uniforme où la probabilité de réponse ne dépend pas de l'information auxiliaire.

Approche inférentielle et les méthodes de repondération

Méthodes de repondération

- Les méthodes de repondération sont généralement utilisées pour composer la non-réponse totale.
- Seulement les données provenant de la réponse totale sont conservées.
- La repondération consiste à hausser les poids de sondage des unités répondantes pour tenir compte des unités non-répondantes.
- Il existe deux méthodes de correction de la non-réponse par repondération : la repondération par l'inverse de la probabilité de réponse et la repondération par calage.
- L'estimateur par repondération du total d'une population est donné par

$$\widehat{t}_y^* = \sum_{k \in s_r} w_k^* y_k$$

où w_k^* est le nouveau poids obtenu avec la méthode de repondération adoptée.

Repondération par l'inverse de la probabilité de réponse

- Dans les expressions des estimateurs des paramètres de la population, au lieu des poids de sondage $w_k = \frac{1}{\pi_k}$, on utilise des poids de sondage ajustés w_k^* définit par :

$$w_k^* = w_k \frac{1}{\widehat{p}_k} = \frac{1}{\pi_k \widehat{p}_k}$$

où $\widehat{p}_k$ est l'estimation de p_k la probabilité de répondre à l'enquête.

- L'estimateur par repondération du total d'une population devient

$$\widehat{t}_y^* = \sum_{k \in s_r} \frac{y_k}{\pi_k \widehat{p}_k}$$

$\hookrightarrow$ Quand $\widehat{p}_k = p_k$, on peut montrer que $\widehat{t}_y^*$ est un estimateur sans biais de t_y.

Repondération par l'inverse de la probabilité de réponse

- Les probabilités de réponse p_k qui sont inconnues varient en général avec les caractéristiques des individus et nécessitent d'émettre des hypothèses sur leur comportement de réponse pour pouvoir estimer les p_k.
- La validité de l'approche inférentielle adoptée dépend de la validité du modèle de comportement de réponse choisi pour estimer les probabilités de réponse p_k.
- Pour estimer les probabilités p_k, le modèle de comportement de réponse le plus utilisé est le modèle logistique définit par

$$p_k = \frac{1}{1 + \exp^{-\mathbf{z}_k' \beta}}$$

où $\mathbf{z}_k = (z_{1k}, \ldots, z_{pk})'$ est le vecteur des valeurs des variables auxiliaires disponibles pour l'unité $k \in s$.

Repondération par l'inverse de la probabilité de réponse

- Souvent on construit des classes de repondération où l'échantillon s est partitionné en H classes $s_1, \ldots, s_h, \ldots, s_H$.
- Ainsi, l'estimateur par repondération de t_y utilisant les H classes de repondération est donné par

$$\widehat{t}^{*}_{y_H} = \sum_{k \in s_r} \frac{w_k y_k}{\widehat{p}_k} = \sum_{h=1}^{H} \frac{1}{\widehat{p}_h} \sum_{k \in s_h} a_k w_k y_k$$

où $a_k = 1$ si $k \in s_r$ et 0 sinon,

$$\overline{y}_{rh} = \frac{\sum_{k \in s_h} w_k a_k y_k}{\sum_{k \in s_h} w_k a_k}$$

Repondération par l'inverse de la probabilité de réponse

- On peut montrer que le biais de $\widehat{t}^{*}_{y_H}$ est donné par

$$\begin{aligned} Biais\left(\widehat{t}^{*}_{y_H}\right) &= E_r\left(\widehat{t}^{*}_{y_H} - \widehat{t}_y \mid s\right) \\ &= \sum_{h=1}^{H} \frac{1}{\overline{p}_h} \sum_{k \in s_h} w_k \left(p_k - \overline{p}_h\right)\left(y_k - \overline{y}_h\right) \end{aligned}$$

où

$$\overline{p}_h = \frac{\sum_{k \in s_h} w_k p_k}{\sum_{k \in s_h} w_k} \quad \text{et} \quad \overline{y}_h = \frac{\sum_{k \in s_h} w_k y_k}{\sum_{k \in s_h} w_k}$$

Repondération par l'inverse de la probabilité de réponse

- Ainsi, le biais de $\widehat{t}^{*}_{y_H}$ est égal à zéro lorsque la covariance entre la probabilité de réponse et la variable d'intérêt est nulle dans chacune des classes.

  cette covariance est nul, par exemple, lorsque le mécanisme de réponse est uniforme à l'intérieur des classes, c'est-à-dire $p_k = p_h$ pour $k \in s_h$

  $\hookrightarrow$ Ce qui correspond à l'hypothèse de l'inférence basée sur le plan.

- En pratique, il est possible de satisfaire cette exigence en construisant des classes de repondération qui soient homogènes par rapport aux probabilités de réponse.

Repondération par calage

- En présence de la non-réponse totale, on fait souvent recours à la repondération par calage pour la corriger.
- Deux procédures de repondération par calage sont utilisées en pratique :
  - ⋆ La procédure à deux étapes :
    - ↪ Première étape : les poids sont modifiés pour corriger la non-réponse (modélisation de la probabilité de réponse).
    - ↪ Deuxième étape : les poids sont de nouveau ajustés de manière à refléter les totaux connus dans la population d'un ensemble de variables auxiliaires.
  - ⋆ La procédure à une seule étape en utilisant le calage généralisé.
- Chacune de ces deux procédures a des avantages et des inconvénients et le choix entre elles n'est pas évident.
  - ↪ Peut-on élaborer un critère permettant de faire ce choix ?

Repondération par calage

- Soit $\widehat{t}^*_{yw}$ l'estimateur obtenu après avoir corrigé la non-réponse en utilisant la repondération par calage :

$$\widehat{t}^*_{yw} = \sum_{k \in s} w_k a_k y_k$$

où $a_k = 1$ si $k \in s_r$ et 0 sinon,

- Pour pouvoir comparer les deux procédures de repondération par calage, on peut se baser sur l'EQM de $\widehat{t}^*_{yw}$ dont l'expression dépend de l'approche inféretielle adoptée.
- Comme, il est difficile de déterminer l'expression de l'EQM de $\widehat{t}^*_{yw}$ sous l'approche basée sur le plan, on peut considérer l'approche basée sur le modèle en ayant pour objectif l'élaboration d'une approximation de

$$EQM_{p_r\xi}(\widehat{t}^*_{yw}) = E_p E_r E_\xi (\widehat{t}^*_{yw} - t_y)^2$$

Repondération par calage

Pour cela, supposons qu'on dispose de

- un ensemble de p variables auxilaires $X_1, \ldots, X_p$ dont les totaux

$$t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$$

sont connus, où $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kp}\right)'$ pour tout $k \in U$.

- un ensemble de p variables instrumentales $Z_1, \ldots, Z_p$ dont le vecteur des valeurs

$$\mathbf{z}_k = \left(z_{k1}, \ldots, z_{kp}\right)' \quad \text{est disponible pour} \quad k \in s_r$$

Les variables instrumentales sont supposées expliquer les probabilités de répondre à l'enquête des unités.

Repondération par calage

Dans le cas de la repondération par calage, les poids de l'estimateur $\widehat{t}^{*}_{yw}$ satisfaient les équations de calage

$$\sum_{k\in S} w_k a_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k$$

où les poids w_k sont censés tenir compte de la non-réponse :

- ⋆ Pour la procédure à deux étapes, les poids de $\widehat{t}^{*}_{yw}$ sont donnés par

  $$w_k = w^{\star}_k \widehat{p}_k^{-1}$$

  où la probabilité de réponse $\widehat{p}_k$ est estimée à l'aide d'un modèle de réponse basé sur les variables instrumentales $Z_1, \ldots, Z_p$.

- ⋆ Pour la procédure à une étape utilisant le calage généralisé, la forme générale des poids de $\widehat{t}^{*}_{yw}$ est donnée par

  $$w_k = d_k F(\boldsymbol{\lambda}' \mathbf{z}_k) \text{ avec } d_k = \frac{1}{\pi_k}$$

  où $F(\cdot)$ est une fonction monotone deux fois différentiable.

Repondération par calage

- Pour l'approche basée sur le modèle, on suppose que les valeurs de $\mathbf{y}$ sont les réalisations d'un modèle de superpopulation ξ donné par

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \epsilon_k$$

avec

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)',$$

$$E_\xi(\epsilon_k) = 0, \quad var_\xi(\epsilon_k) = \sigma^2 v_k^2 \text{ et } cov_\xi(\epsilon_k, \epsilon_l) = 0.$$

Les v_k^2 sont supposé connus avec $\sum_{k \in U} v_k = N$

- De plus, sous l'approche basée sur le plan et le modèle et en présence de la non-réponse, la précision de $\widehat{t}^*_{yw}$ peut être mesurée par

$$EQM_{p_r\xi}(\widehat{t}^*_{yw}) = E_p E_r E_\xi (\widehat{t}^*_{yw} - t_y)^2$$

où E_p, E_r et E_ξ sont respectivement les espérances sous le plan, le mécanise de non-réponse et le modèle de superpopulation.

Repondération par calage

Approximation de $EQM_{p_r\xi}$

Sous l'approche basée sur le plan et le modèle, une approximation de la précision de l'estimateur $\widehat{t}^*_{yw}$ peut être donnée par

$$EQM_{p_r\xi}(\widehat{t}^*_{yw}) \approx \sigma^2 \sum_{k\in U} v_k^2 \left[(d_k - 1)\frac{(p_k w_k)^2}{d_k^2} + \left[\frac{p_k w_k}{d_k} - 1\right]^2 \right]$$

où d_k *est le poids de sondage,*
w_k *est le poids de calage,*
p_k *est la probabilité de réponse.*

Repondération par calage

La comparaison entre l'estimateur par calage généralisé et la procédure à deux étapes peut être réalisée en utilisant l'estimateur de l'approximation de $EQM_{p_r\xi}(\widehat{t}^*_{yw})$ donnée par

$$\widehat{EQM}_{p_r\xi}(\widehat{t}^*_{yw}) = \widehat{\sigma^2}\sum_{k\in S} d_k v_k^2 \left[(d_k - 1)\frac{(\widehat{p}_k w_k)^2}{d_k^2} + \left[\frac{\widehat{p}_k w_k}{d_k} - 1\right]^2\right]$$

où $\widehat{\sigma^2}$ est l'estimateur de la variance des résidus

- de la régression de Y en fonction de $X_1, \ldots, X_p$, dans le cas de la procédure à deux étapes avec $\widehat{p}_k$ est l'estimateur de p_k obtenu à la première étape.
- de la régression de Y en fonction régression instrumentale de Y en fonction de $X_1, \ldots, X_p$ utilisant $Z_1, \ldots, Z_p$ comme variables instrumentales, dans le cas de l'estimateur par calage généralisé avec $\widehat{p}_k^{-1} = F(\boldsymbol{\lambda}^{'}\mathbf{z}_k)$.

Repondération par calage

Notons que dans le cas de l'estimateur par calage généralisé ($\widehat{p}_k^{-1} = F(\boldsymbol{\lambda}^{'}\mathbf{z}_k)$), on a

$$\widehat{EQM}_{p_r\xi}(\widehat{t}_{yw}) = \widehat{\sigma^2}\sum_{k\in U} v_k^2(d_k - 1)$$

avec $\widehat{\sigma^2}$ est l'estimateur de la variance des résidus de la régression instrumentale de Y en fonction de $X_1, \ldots, X_p$ utilisant $Z_1, \ldots, Z_p$ comme variables instrumentales.

- $\hookrightarrow$ Comparé à la procédure à deux étapes, l'utilisation du calage généralisé permet de réduire le biais dû à la non-réponse mais ceci peut être au détriment d'une perte en termes de variance.
- $\hookrightarrow$ Si le gain en termes de réduction du biais de non-réponse dépasse la perte éventuelle en termes d'augmentation de la variance, l'estimateur par calage généralisé est préféré à celui utilisant la procédure à deux étapes.

Repondération par calage

- L'utilisation de l'approche modèle a permis d'élaborer un critère du choix entre les deux procédures de correction de non-réponse utilisant la repondération par calage suivants :
  - ⋆ La procédure à deux étapes.
  - ⋆ La procédure à une seule étape en utilisant le calage généralisé.
- Ce critère a l'avantage de tenir compte de l'impact de l'utilisation du calage sur le biais et la variance de l'estimateur obtenu après correction de la non-réponse en utilisant la repondération par calage.

Plan

Problèmatique dans les petits domaines

- La précision des estimateurs dépend de la taille de l'échantillon : lorsque cette taille est faible, il y a un fort risque que la qualité de la précision des estimateurs soit faible.
- Cette situation peut être rencontrée lorsqu'on s'intéresse aux sous-populations (petits domaines) à partir desquelles peu d'unités ont été sélectionnées dans l'échantillon.
- C'est pour cela, l'estimation des paramètres pour les petits domaines fait recours à des techniques différentes de celles utilisées lorsqu'il s'agit de la population entière.
- L'objectif des techniques d'estimation sur petits domaines est d'essayer de corriger les erreurs dues au fait que la taille de domaine est petite.

Problèmatique dans les petits domaines

- On dispose d'une population $U = \{1, \ldots, k, \ldots, N\}$ de taille N.
- Un domaine U_d composé de N_d unités.
- Une variable d'intérêt $Y : U \to \mathbb{R}$ dont les valeurs sont notées par $y_1, \ldots, y_k, \ldots, y_N$.
- Un échantillon s de n unités est sélectionné à partir de U selon un plan de sondage $p(s)$ dont les probabilités d'inclusion sont données par $\pi_k = P(k \in s)$ et $\pi_{kl} = P(k, l \in s)$
- **Objectif** : On s'intéresse à l'estimation d'une fonction des valeurs de la variable d'intérêt spécifique au domaine U_d :

$$t_{yd} = \sum_{k \in U_d} y_k \quad \text{et} \quad \mu_{yd} = \frac{1}{N_d} \sum_{k \in U_d} y_k$$

Catégories d'estimateurs pour les petits domaines

Pour estimer les paramètres d'intérêt relatifs à un petit domaine, on distingue trois catégories d'estimateurs :

(1) Estimateurs directs :

- Ce type d'estimateurs n'utilise aucune information hors du domaine.
- On peut utiliser l'information auxiliaire mais en se limitant à celle disponible au sein du domaine.
- Ces estimateurs sont faciles à utiliser mais ils sont en général moins précis.
- Il est préférable de les utiliser lorsque la taille du domaine n'est pas petite ou lorsque la construction du domaine a été prise en compte par le plan de sondage.

Catégories d'estimateurs pour les petits domaines

(2) Estimateurs indirects avec modélisation implicite :

- Ces estimateurs utilisent un modèle dont l'objectif est d'expliquer les caractéristiques de la variable d'intérêt Y tout en permettant de relier les unités du domaine à celle de la population U.
- Notons que le modèle utilisé porte sur des hypothèses relatives aux données agrégées. Par exemple, on suppose que la moyenne de la variable d'intérêt Y dans le domaine est égale à celle dans la population. La précision de ces estimateurs dépend de la validité de ces hypothèses.
- Il n'y a pas d'autre aléa que celui de l'échantillonnage (approche basée uniquement sur le plan).

Catégories d'estimateurs pour les petits domaines

(3) Estimateurs indirects avec modélisation explicite :

- Ces estimateurs se basent sur un modèle utilisant des variables auxiliaires explicatives pour expliquer les caractéristiques de la variable d'intérêt Y et une composante stochastique.
- Ces estimateurs sont plus complexes à mettre en oeuvre mais ils sont en général plus précis surtout en présence d'information auxiliaire assez riche et actualisée.
- Pour les modèles utilisés, l'unité statistique peut être :
  - Le petit domaine : lorque des estimateurs (directs) sont disponibles pour plusieurs domaines et en présence d'information auxiliaire au niveau de ces domaines.
  - L'individu de la population : dans le cas où l'information auxiliaire est disponible au niveau individu.

Catégories d'estimateurs pour les petits domaines

A travers ces différents catégories d'estimateurs utilisés pour l'estimation sur les petits domaines, on note que :

- Comme les estimateurs directs sont moins précis, une amélioration de la qualité des estimations sur petits domaines peut être obtenue en utilisant l'information auxiliaire disponible en dehors du domaine.
- Le lien entre le domaine et le reste de la population ne peut se faire qu'à travers la supposition d'un modèle de superpopulation.
- L'utilisation de l'approche modèle en estimation sur petits domaines a permis de proposer des estimateurs qui peuvent fournir des estimations plus précises en particulier en présence d'information auxiliaire riche et lorsque le modèle supposé est bien spécifié.

Exemple de méthodes d'estimation sur petits domaines sous l'approche modèle

- Sous l'approche modèle, les estimateurs des paramètres de la population se caractérisent par une distribution de probabilité issue d'une double experience aléatoire :
  - La disribution du modèle de superpopulation ξ qui fournit la population U d'intérêt.
  - La disribution du plan de sondage $p(s)$ utilisé pour sélectionner l'échantillon s.

- Pour l'estimation sur petits domaines, la classe des estimateurs sans biais optimums et linéaires (BLUE) est considérée parmi celle les plus utilisées en pratique.
- Les estimateurs BLUE s'appliquent dans le cas des petits domaines en présence de la classe de Modèles Linéaires Mixtes : il s'agit de modèles expliquant une variable d'intérêt Y à l'aide d'effets fixes (de type régression linéaire) et d'effets aléatoires.

Modèles linéaires mixtes

La formulation générale du modèle linéaire mixte est la suivante :

$$\mathbf{Y}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{Z}_s\boldsymbol{v} + \mathbf{e}_s$$

Où

- $\mathbf{Y}_s$ le vecteur des n valeurs observées de la variable d'intérêt Y
- $\mathbf{X}_s$ matrice $n \times J$ des variables auxiliaires (connue)
- $\boldsymbol{\beta}$ vecteur de J coefficients de régression $\longrightarrow$ **effets fixes**
- $\mathbf{Z}_s$ matrice $n \times p$ des variables auxiliaires (connue)
- $\boldsymbol{v}$ vecteur aléatoire de taille p $\longrightarrow$ **effets aléatoires**
- $\mathbf{e}_s$ vecteur aléatoire de taille n.

Modèles linéaires mixtes

De plus, le modèle linéaire mixte suppose que :

- $E_\xi(\boldsymbol{v}) = 0$ et $E_\xi(\mathbf{e}) = 0$.
- $\mathbf{e}$ et $\boldsymbol{v}$ sont indépendants.
- $Var_\xi(\boldsymbol{v}) = \mathbf{G}(\boldsymbol{\delta})$ et $Var_\xi(\boldsymbol{e}_s) = \mathbf{R}_s(\boldsymbol{\delta})$
  où $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_q)$ est un vecteur de paramètres inconnus.
- Pour les variances $\mathbf{G}(\boldsymbol{\delta})$ et $\mathbf{R}_s(\boldsymbol{\delta})$, leur expressions sont connues mais elles ne peuvent être calculées car $\boldsymbol{\delta}$ est inconnu.

Estimateur sans biais linéaire optimal (BLUE)

- Supposons qu'on cherche à estimer la valeur réelle de

$$\mu = \mathbf{h}^{'}\boldsymbol{\beta} + \mathbf{m}^{'}\boldsymbol{v}$$

où $\mathbf{h}^{'}$ et $\mathbf{m}^{'}$ sont des vecteurs connus.

- Pour estimer μ, on va se restreindre par souci de simplicité aux estimateurs linéaires de type

$$\widehat{\mu} = \mathbf{a}^{'}\mathbf{Y}_s + b$$

où $\mathbf{a}$ et b sont connus.

- De plus, on va imposer à l'estimateur $\widehat{\mu}$ d'être sans biais au sens suivant

$$E_\xi(\widehat{\mu} - \mu) = 0$$

Estimateur sans biais linéaire optimal (BLUE)

Ainsi, suite à la recherche de l'estimateur sans biais optimal au sens de la minimisation de l'Erreur Quadratique Moyen sous le modèle :

$$E_\xi(\widehat{\mu}-\mu)^2$$

l'estimateur sans biais linéaire optimal (BLUE) de μ est donné par :

$$\widehat{\mu}_{BLUE} = \mathbf{a}^{'}_{BLUE}\mathbf{Y}_s + b_{BLUE}$$

où

$$b_{BLUE} = 0$$

$$\mathbf{a}^{'}_{BLUE} = \left(\mathbf{h}^{'} - \mathbf{m}^{'}\mathbf{G}\mathbf{Z}_s^{'}\mathbf{V}_s^{-1}\mathbf{X}_s\right)\left(\mathbf{X}_s^{'}\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s^{'}\mathbf{V}_s^{-1} + \mathbf{m}^{'}\mathbf{G}\mathbf{Z}_s^{'}\mathbf{V}_s^{-1}$$

avec

$$\mathbf{V}_s = Var_\xi\left(\mathbf{Y}_s\right) = Var_\xi\left(\mathbf{Z}_s\boldsymbol{v}\right) + Var_\xi\left(\mathbf{e}_s\right) = \mathbf{Z_s}\mathbf{G}\mathbf{Z}_s^{'} + \mathbf{R}_s$$

Estimateur sans biais linéaire optimal (BLUE)

- Afin de faciliter l'interprétation de l'expression de $\widehat{\mu}_{BLUE}$, il veut mieux la réécrire comme suit :

$$\widehat{\mu}_{BLUE} = \mathbf{h}^{'}\widetilde{\boldsymbol{\beta}} + \mathbf{m}^{'}\left[\mathbf{G}\mathbf{Z}_{s}^{'}\mathbf{V}_{s}^{-1}\left(\mathbf{Y}_{s} - \mathbf{X}_{s}\widetilde{\boldsymbol{\beta}}\right)\right]$$

où $\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X}_{s}^{'}\mathbf{V}_{s}^{-1}\mathbf{X}_{s}\right)^{-1}\mathbf{X}_{s}^{'}\mathbf{V}_{s}^{-1}\mathbf{Y}_{s}$

- On note que l'expression de $\widetilde{\boldsymbol{\beta}}$ correspond à celle du vecteur des coefficients de régression $\boldsymbol{\beta}$ obtenue avec la méthode des moindres carrés généralisés pour le modèle de régression $\mathbf{Y}_{s} = \mathbf{X}_{s}^{'}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{s}$
  - $\hookrightarrow$ On peut donc interpréter le second terme dans l'expression de $\widehat{\mu}_{BLUE}$ comme le meilleur estimateur de la variable aléatoire $\boldsymbol{v}$
- On note également qu'on ne peut pas calculer $\mathbf{a}_{BLUE}$ et par conséquent $\widehat{\mu}_{BLUE}$ que si $\boldsymbol{\delta}$ est connu. C'est pour cela qu'il faut l'estimer car en général il est inconnu.

Estimateur sans biais linéaire optimal (BLUE)

- Pour mesurer la précision de l'estimateur $\widehat{\mu}_{BLUE}$, on peut utiliser l'Ecart Quadratique Moyen (EQM) :

$$EQM(\widehat{\mu}_{BLUE}) = E_{\xi}\left(\widehat{\mu}_{BLUE} - \mu\right)^2 = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta})$$

où

$$g_1(\boldsymbol{\delta}) = \mathbf{m}^{'}\left(\mathbf{G} - \mathbf{G}\mathbf{Z}_s^{'}\mathbf{V}_s^{-1}\mathbf{Z}_s\mathbf{G}\right)\mathbf{m}$$

$$g_2(\boldsymbol{\delta}) = \left(\mathbf{h}^{'} - \mathbf{m}^{'}\mathbf{G}\mathbf{Z}_s^{'}\mathbf{V}_s^{-1}\mathbf{X}_s\right)\left(\mathbf{X}_s^{'}\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\left(\mathbf{h} - \mathbf{X}_s^{'}\mathbf{V}_s^{-1}\mathbf{Z}_s\mathbf{G}^{'}\mathbf{m}\right)$$

- L'expression de $EQM(\widehat{\mu}_{BLUE})$ est complexe mais on peut la calculer lorsque $\boldsymbol{\delta}$ est connu.

Estimateur sans biais linéaire optimal (BLUE)

- Le calcul de l'estimateur $\widehat{\mu}_{BLUE}$ ne peut se faire qu'après avoir calculé les valeurs des éléments des matrices de variance pour $\boldsymbol{v}$ et $\mathbf{e}$ à travers la connaissance des valeurs du vecteur des paramètres $\boldsymbol{\delta}$.
- Malheureusement, les valeurs du vecteur des paramètres $\boldsymbol{\delta}$ sont rarement connues, ce qui oblige les statisticiens de l'estimer en utilisant les informations auxiliaires disponibles sur la variable d'intérêt Y.
- Pour estimer $\boldsymbol{\delta}$, on peut utiliser la méthode de Maximum de vraissemblance :
  - $\hookrightarrow$ On peut utiliser la procédure PROC MIXED contenue dans le logiciel SAS $\rightarrow$ Estimation de $\boldsymbol{\delta}$ et de $\boldsymbol{\beta}$.

Estimateur BLUE avec le modèle de Fay et Herriot

Le modèle de Fay et Herriot est parmi les modèles conçu au niveau des domaines. En effet, si on s'intéresse à l'estimation d'un paramètre $\theta_d = g(t_{yd})$ relatif au domaine d $(d = 1, \ldots, m)$, l'expression du modèle de Fay et Herriot est la suivante :

$$\widehat{\theta}_d = \mathbf{z}_d^{'}\boldsymbol{\beta} + b_d v_d + e_d$$

où

- e_d $(d = 1, \ldots, m)$ sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) avec $E_\xi(e_d) = 0$ et $Var_\xi(e_d) = \Psi_d$ où Ψ_d est connu.
- Les variables e_d et v_d sont indépendantes.
- La partie aléatoire du modèle $b_d v_d + e_d$ a une espérance nulle et une variance égale à $b_d^2\sigma_v^2 + \Psi_d$.

Estimateur BLUE avec le modèle de Fay et Herriot

- Comme le modèle de Fay et Herriot vise à expliquer la variable d'intérêt Y en considérant un terme fixe et un autre alétoire, il est considéré parmi les modèles mixtes conçus au niveau des domaines pour lesquels on s'intéresse à estimer pour chaque domaine d,

$$\mu_d = \mathbf{h}_d^{'}\boldsymbol{\beta} + \mathbf{m}_d \boldsymbol{v}_d \qquad \Rightarrow \qquad \theta_d = \mathbf{z}_d^{'}\boldsymbol{\beta} + b_d v_d$$

avec $\mu_d = \theta_d$, $\mathbf{h}_d = \mathbf{z}_d$, $\mathbf{m}_d = b_d$ et $\boldsymbol{v}_d = v_d$.

- En effet, à travers l'expression générale des modèles mixtes

$$\mathbf{Y}_{s_d} = \mathbf{X}_{s_d}\boldsymbol{\beta} + \mathbf{Z}_{s_d}\boldsymbol{v}_d + \mathbf{e}_{s_d}$$

on a,

$$\widehat{\theta}_d = \mathbf{z}_d^{'}\boldsymbol{\beta} + b_d v_d + e_d$$

avec $\mathbf{Y}_{s_d} = \widehat{\theta}_d$, $\mathbf{X}_{s_d} = \mathbf{z}_d$, $\mathbf{Z}_{s_d} = b_d$ et $\boldsymbol{v}_d = v_d$.

Estimateur BLUE avec le modèle de Fay et Herriot

L'expression de la variance sous le modèle de Fay et Herriot de l'estimateur BLUE peut être déduite à partir de celle obtenue sous les modèles mixtes où pour chaque domaine d, on a

$$\mathbf{V}_{s_d} = \mathbf{Z}_{s_d}\mathbf{G}_d\mathbf{Z}'_{s_d} + \mathbf{R}_{s_d}$$

avec $\mathbf{Z}_{s_d} = b_d$, $\mathbf{G}_d = \sigma_v^2$ et $\mathbf{R}_{s_d} = \Psi_d$.

Ainsi, la variance de $\widehat{\theta}_d$ sous le modèle de Fay et Herriot est donnée par :

$$V\left(\widehat{\theta}_d\right) = b_d^2\sigma_v^2 + \Psi_d$$

Estimateur BLUE avec le modèle de Fay et Herriot

A partir de l'expression générale de l'estimateur $\widehat{\mu}_{d,BLUE}$ sous l'hypothèse du modèle linéaire mixte :

$$\widehat{\mu}_{d,BLUE} = \mathbf{h}_d^{'}\widetilde{\boldsymbol{\beta}} + \mathbf{m}_d^{'}\left[\mathbf{G}_d\mathbf{Z}_{s_d}^{'}\mathbf{V}_{s_d}^{-1}\left(\mathbf{Y}_{s_d} - \mathbf{X}_{s_d}\widetilde{\boldsymbol{\beta}}\right)\right]$$

On peut déduire celle de l'estimateur $BLUE$ de θ_d sous l'hypothèse du modèle de Fay et Herriot qui est donnée par :

$$\widehat{\theta}_{d,BLUE} = \mathbf{z}_d^{'}\widetilde{\boldsymbol{\beta}} + \Gamma_d\left(\widehat{\theta}_d - \mathbf{z}_d^{'}\widetilde{\boldsymbol{\beta}}\right)$$

où

$$\Gamma_d = \frac{b_d^2\sigma_v^2}{\Psi_d + b_d^2\sigma_v^2}$$

et

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{m}\frac{\mathbf{z}_d\mathbf{z}_d^{'}}{\Psi_d + b_d^2\sigma_v^2}\right)^{-1}\sum_{d=1}^{m}\frac{\mathbf{z}_d\widehat{\theta}_d}{\Psi_d + b_d^2\sigma_v^2}$$

Estimateur BLUE avec le modèle de Fay et Herriot

- Le calcul de l'estimateur $\widehat{\theta}_{d,BLUE}$ nécessite l'estimation de la variance σ_v^2 qui peut se faire en utilisant la méthode des Moments ou celle du Maximum de vraissemblance.
- On note que pour l'estimateur $\widehat{\theta}_{d,BLUE}$, l'expression de $\widetilde{\boldsymbol{\beta}}$ prend en compte tous les domaines, ce qui le rend plus stable :

$$V\left(\widetilde{\boldsymbol{\beta}}\right) = O\left(\frac{1}{\sum_{d=1}^{m} n_d}\right)$$

Estimateur BLUE avec le modèle de Fay et Herriot

On note qu'on peut réécrire l'expression de l'estimateur $\widehat{\theta}_{d,BLUE}$ comme suit

$$\widehat{\theta}_{d,BLUE} = \Gamma_d \underbrace{\widehat{\theta}_d}_{\text{Estimateur direct}} + (1-\Gamma_d) \underbrace{\mathbf{z}_d'\widetilde{\boldsymbol{\beta}}}_{\text{Estimateur synthétique}}$$

* $\Gamma_d \in [0,1] \quad \Rightarrow \quad$ l'estimateur $\widehat{\theta}_{d,BLUE}$ est un estimateur composite de θ_d
* Quand les valeurs de b_d et σ_v^2 sont faibles (l'effet de v_d est faible), $\widehat{\theta}_{d,BLUE}$ devient presque égal à l'estimateur synthétique.
* Quand la variance Ψ_d est faible, on a $\Gamma_d \approx 1$ et $\widehat{\theta}_{d,BLUE}$ devient presque égal à l'estimateur direct.

Estimateur BLUE avec le modèle de Fay et Herriot

* Lorsque $\sigma_v^2 = 0$, on a $\Gamma_d = 0$ et par conséquent, l'estimateur $\widehat{\theta}_{d,BLUE}$ devient presque égal à l'estimateur synthétique.

  $\Rightarrow$ Quand il n'y a pas d'effet spécifique pour chaque domaine, l'estimateur synthétique est le meilleur parmi les estimateurs linéaires sans biais.

* Pour estimer θ_d, on peut considérer l'estimateur $\widehat{\theta}_{d,BLUE}$ même lorsque $n_d = 0$ et ceci en utilisant le terme synthétique de $\widehat{\theta}_{d,BLUE}$.

Evaluation de la qualité de l'estimateur BLUE

L'estimateur $\widehat{\theta}_{d,BLUE}$ possède les proptiétés suivantes :

* **Propriété 1** Quand $n_d \longrightarrow N_d$, on a $\Psi_d \longrightarrow 0$ et $\Gamma_d \longrightarrow 1$. Dans ce cas, on a

$$\widehat{\theta}_{d,BLUE} \approx \widehat{\theta}_d \approx \theta_d$$

* **Propriété 2** Quand on se limite à la distribution du plan de sondage, $\widehat{\theta}_{d,BLUE}$ est un estimateur biaisé de θ_d et on a

$$E_p\left(\widehat{\theta}_{d,BLUE}\right) - \theta_d = (1-\Gamma_d)\left[\mathbf{z}_d' E_p\left(\widetilde{\boldsymbol{\beta}}\right) - \theta_d\right]$$

$\widehat{\theta}_{d,BLUE}$ ne devient sans biais que lorsqu'on tient en compte la distribution du modèle de superpopulation :

$$E_\xi\left[E_p\left(\widehat{\theta}_{d,BLUE}\right) - \theta_d\right] = 0$$

Evaluation de la qualité de l'estimateur BLUE

* **Propriété 3** Pour l'estimateur $\widehat{\theta}_{d,BLUE}$, on peut montrer que

$$\begin{aligned} EQM\left(\widehat{\theta}_{d,BLUE}\right) &= \Gamma_d \Psi_d + (1-\Gamma_d)^2 \mathbf{z}_d' \left(\sum_{d=1}^{m} \frac{\mathbf{z}_d \mathbf{z}_d'}{\Psi_d + \sigma_v^2 b_d^2}\right)^{-1} \mathbf{z}_d \\ &= \Gamma_d \Psi_d + O\left(\frac{1}{m}\right) \\ &\approx \Gamma_d \Psi_d \qquad \text{quand } m \text{ est grand} \end{aligned}$$

$$\Rightarrow \qquad \Gamma_d \approx \frac{EQM\left(\widehat{\theta}_{d,BLUE}\right)}{EQM\left(\widehat{\theta}_d\right)}$$

Le gain en précision obtenu en utilisant l'estimateur $\widehat{\theta}_{d,BLUE}$ augmente quand la valeur de Γ_d est faible.

Estimateur BLUE avec le modèle de Fay et Herriot

On note qu'en pratique, l'expression utilisée de $\widehat{\theta}_{d,BLUE}$ est celle obtenue après avoir remplacé σ_v^2 par son estimateur $\widehat{\sigma}_v^2$, et qui est donnée par

$$\widehat{\theta}_{d,BLUE} = \widehat{\Gamma}_d \widehat{\theta}_d + \left(1 - \widehat{\Gamma}_d\right) \mathbf{z}_d' \widehat{\boldsymbol{\beta}}$$

où

$$\widehat{\Gamma}_d = \frac{b_d^2 \widehat{\sigma}_v^2}{\Psi_d + b_d^2 \widehat{\sigma}_v^2}$$

et

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{m} \frac{\mathbf{z}_d \mathbf{z}_d'}{\Psi_d + b_d^2 \widehat{\sigma}_v^2} \right)^{-1} \sum_{d=1}^{m} \frac{\mathbf{z}_d \widehat{\theta}_d}{\Psi_d + b_d^2 \widehat{\sigma}_v^2}$$

Estimateur BLUE avec le modèle de Fay et Herriot

A travers l'expression de l'estimateur $\widehat{\theta}_{d,EBLUP}$, on peut considérer une procédure d'estimation qui est un compromis entre l'approche basée sur le plan et celle basée sur le modèle tout en favorisant l'approche la plus fiable.
Pour cela, Fay et Herriot propose d'utiliser l'estimateur suivant :

$$\widehat{\theta}^{*}_{d,BLUE} = \begin{cases} \widehat{\theta}_{d,BLUE} & \text{Si} \quad \mid \widehat{\theta}_{d} - \widehat{\theta}_{d,BLUE} \mid \leq c\sqrt{\Psi_{d}} \\ \widehat{\theta}_{d,BLUE} - c\sqrt{\Psi_{d}} & \text{Si} \quad \widehat{\theta}_{d,BLUE} < \widehat{\theta}_{d} - c\sqrt{\Psi_{d}} \\ \widehat{\theta}_{d,BLUE} + c\sqrt{\Psi_{d}} & \text{Si} \quad \widehat{\theta}_{d,BLUE} > \widehat{\theta}_{d} + c\sqrt{\Psi_{d}} \end{cases}$$

$\Rightarrow$ La variance de l'estimateur $\widehat{\theta}^{*}_{d,BLUE}$ est plus faible de celles des estimateurs $\widehat{\theta}_{d,BLUE}$ et $\widehat{\theta}_{d}$.

Remarques et conclusion

- Pour les estimations sur petits domaines, la mesure de la qualité des estimations obtenues ne peut se faire sans recours à des hyppothèses.
- Par exemple, on ne pourra pas tenir compte des spécificités des domaines qui n'ont pas été prises en compte au travers d'un modèle sans faire des hypothèses.
- Pour quantifier la pertinence de l'estimation, il n'existe pas de méthodes infaillibles pour qualifier la pertinence de l'estimation mais néanmoins, on dispose des critères d'appréciation de la qualité de celle-ci.

Remarques et conclusion

Les principaux critères utilisés pour apprécier la qualité de l'estimation obtenue sont les suivants :

(1) Vérifier la qualité du modèle sur lequel se base les estimateurs utilisés :

- Choix des variables explicatives qui convient d'utiliser dans le modèle.
- L'étude du graphique des résidus estimés du modèle qui permettent de vérifier l'hypothèse d'espérance nulle pour ces résidus ainsi que d'apprécier la validité de l'hypothèse formulée sur leur variance.
- Le calcul des coefficients qui mesurent la qualité d'ajustement du modèle aux données observées comme par exemple le coefficient R^2.
- La détection des individus ayant une influence particulièrement forte dans la détermination des paramètres du modèle, ce qui permet ensuite de les traiter de manière spécifique.

Remarques et conclusion

(2) Le calcul de l'Erreur Quadratique Moyen (EQM) des estimateurs sur petits domaines :

↪ En effet, au travers l'EQM, l'approche modèle permet de produire des outils intéressants pour mesurer la qualité des estimations.

On note que la fiabilité de l'EQM augmente lorsque le modèle tient compte des spécificités des petits domaines.

(3) Dans le cas où on dispose des domaines dont les tailles sont jugées suffisament grandes, ces domaines peuvent être considérés pour évaluer la pertinence des techniques d'estimations pour petits domaines en comparant les résultats de ces techniques à ceux des estimateurs directs.

Remarques et conclusion

La majorité des méthodes d'estimation sur petits domaines se caractérisent s'appuient sur un modèle. En effet, pour

- La majorité des méthodes d'estimation sur petits domaines se caractérise par le fait qu'elle s'appuie sur un modèle.
- En effet, pour pouvoir produire des estimations fiables pour les petits domaines, on doit mettre à profit les infomations disponibles en dehors de ces domaines.
- C'est pour cela qu'on ne peut pas se passer d'utiliser un modèle dont le rôle est de faire le lien entre le petit domaine et la source d'information disponible en dehors de domaine.
- L'utilisation de l'approche modèle en estimation sur petits domaines a permis d'élaborer des techniques d'estimation plus pertinantes surtout lorsque le modèle adopté est bien spécifié.