

LA CORRECTION DE LA NON-RÉPONSE PAR CALAGE OU PAR ÉCHANTILLONNAGE ÉQUILIBRÉ

Jean-Claude DEVILLE¹

RÉSUMÉ

La correction de la non-réponse totale se fait généralement par une repondération des unités élémentaires. Celle-ci résulte logiquement de l'estimation à partir des données d'un modèle paramétrique de réponse. Les corrections de poids apparaissent alors comme les inverses de probabilités de réponse estimées. Cette façon de voir les choses amène à se poser des questions sur les méthodes d'estimation à mettre en œuvre pour parvenir à ces fins. Il apparaît que l'utilisation d'un principe de calage arrange singulièrement bien les choses en autorisant une réduction de la variance causée par l'aléa de non-réponse, en permettant aussi de la calculer et de l'estimer. Au passage, on est amené à développer une forme de calage un peu plus générale que celle de la théorie habituelle et qui peut présenter un intérêt autonome.

La correction de la non-réponse partielle se fait généralement par imputation. Parmi les différentes méthodes possibles celles qui font appel à des donneurs (hot-deck comme on dit en France) sont particulièrement appelantes à causes de leur simplicité et de leur caractère non-paramétrique. Elles entretiennent aussi des relations assez étonnantes avec la théorie de l'échantillonnage. En particulier le recours à des techniques d'échantillonnage contraint (ou équilibré) permet de diminuer très fortement la variance parasite et inutile introduite par le choix aléatoire des donneurs.

Le présent texte développera essentiellement la première partie, qui a été exposée oralement. On se contentera d'indications rapides pour la deuxième partie.

MOTS-CLÉS : Non-réponse totale; calage; échantillonnage équilibré.

ABSTRACT

The complete non response correction is generally made by reweighting the elementary units. It results naturally from the estimation (using the responses) of a parametric response model. The weight corrections appear to be like the inverse of the estimated response probabilities. In the estimation procedure the use of the benchmarking principle works well since it allows for a variance reduction due to the randomness of the non response, its evaluation and its estimation. In doing so, we are led to develop new benchmarking techniques more general than those provided by the general theory. This new development may present an interest of its own.

The partial non response correction is usually made by imputation. Among the different possible methods (hot-deck methods as we call them in France which use donors) are particularly appealing because of their simplicity and their non parametric aspect. They are also astonishingly related to the sampling theory. Particularly the use of constrained sampling techniques allows a very substantial reduction of the noisy and useless variance introduced by the random choice of the donors.

The first part of the following paper covers what was presented at the conference. The second part will only give quick explanations.

KEYS WORDS : Non Response; Calibration; Balanced Sampling.

¹ Institut National de la Statistique et des Etudes Economiques Unités Méthodes Statistiques, 18, Boulevard Adolphe Pinard
- 75675 Paris cédex 14 - jean-claude.deville@insee.fr

1. QUELQUES IDÉES GÉNÉRALES SUR LA CORRECTION POUR NON-RÉPONSE

La compensation de la non-réponse dans les enquêtes s'appuie sur deux techniques différentes : la repondération destinée essentiellement à corriger la non-réponse globale, et l'imputation destinée à corriger la non-réponse partielle.

Les procédures de repondération sont basées sur une modélisation du mécanisme de réponse. Celui-ci consiste en la sélection d'un échantillon r de répondants selon un "pseudo" plan de sondage $q(r|s)$, inconnu. En paramétrant ce pseudo plan à l'aide d'un paramètre β de dimension p on cherchera à estimer β et tout se passera comme si le paramètre estimé était le bon paramètre. C'est pour cela qu'on parlera de méthode de pseudo-randomisation.

Parallèlement, la théorie de l'estimation en sondage a connu récemment des développements intéressants qui synthétisent et généralisent les types d'estimateurs qu'on utilisait traditionnellement (Deville, Särndal (1992), Deville, Särndal, Sautory (1993)). Il existe une analogie formelle (en fait assez profonde) entre ces estimateurs et les modèles linéaires généralisés qui permettent une description intéressante du mécanisme de réponse. Ceci conduit à une synthèse entre ces approches déjà ébauchée dans (Deville, Dupont (1993)) et qui trouvera ici une extension nouvelle basée sur la théorie généralisée du calage. On arrive par cette voie à de nouvelles procédures de repondération qui s'avèrent à la fois commodes et efficaces. En particulier on met en évidence des gains de variance dus à la procédure même d'estimation. Ce sera le principal objet de cet exposé.

Les idées courantes sur l'imputation sont basées sur des modèles permettant de prédire la variable d'intérêt y_k à partir de cofacteurs. Éventuellement on estimera, à partir des données fournies par les répondants, la loi de probabilité suivie par y_k conditionnellement aux cofacteurs. Il y a dès lors deux variantes usuelles :

- Imputer la meilleure prédiction
- Imputer une valeur prise au hasard dans la loi estimée de y_k .

Les deux types de méthodes ont des avantages et des inconvénients. Nous verrons, à la fin de cet article, qu'on peut cependant développer des méthodes mixtes basées sur l'échantillonnage équilibré, qui réalisent un compromis intéressant entre ces deux variantes. Nous

verrons aussi qu'il existe des possibilités de pseudo-pondération basées sur les mêmes idées.

La suite de ce papier est essentiellement consacrée aux méthodes de repondération qui seront analysées en détail et de façon assez générale. Pour en arriver là nous aurons besoin d'une extension des techniques de calage qui seront survolées au paragraphe 2. On passera ensuite à l'exploitation de ces idées dans le cadre d'une enquête exhaustive entachée de non-réponse. On passera ensuite au cas général, puis au cas où une information auxiliaire est présente à deux niveaux : celui de la population et celui de l'échantillon sans non-réponse.

On poursuivra par une indication rapide de ce qu'on peut faire en matière d'imputation.

2. THÉORIE GÉNÉRALISÉE DU CALAGE : QUELQUES IDÉES UTILES

2.1 Que savons-nous ?

Rappelons les lignes générales de la théorie classique du calage. A partir d'un estimateur sans biais, ou, du moins, "convergent", en général celui de Horvitz-Thompson, $\hat{Y} = \sum_s d_k y_k$ du total Y des y_k on cherche de nouveaux poids w_k proches des d_k au sens d'une certaine distance et vérifiant les équations de calage : $X = \sum_s w_k x_k$ où X est le p -vecteur de totaux de p variables réelles rangées dans le vecteur x_k . On trouve alors que les poids sont de la forme $w_k = F(q_k x_k' \lambda)$ où les q_k sont une famille de nombres positifs destinés à prendre en compte une certaine hétéroscédasticité, λ un p -vecteur et F une fonction régulière de \mathbf{R} dans \mathbf{R} vérifiant $F(0) = 1$ et $F'(0) = 1$.

L'intérêt de cette procédure d'estimation est d'éliminer, dans la variance de l'estimateur, l'influence des variables x_k . De façon précise on a la résultat

suivant: $\text{Var}\left(\sum_s w_k y_k\right) = \text{Var}\left(\sum_s w_k e_k\right)$, où les e_k

sont les résidus de la régression $y_k = x_k' \beta + e_k$ obtenues par les moindres carrés pondérés par les nombre q_k .

2.2 Ébauche d'une théorie généralisée.

A chaque k de la population on associe une fonction de calage $F_k : \mathbb{R}^p \rightarrow \mathbb{R}$ vérifiant $F_k(0) = 0$, régulière, et on résout les équations de calage :

$$X = \sum_s d_k x_k F_k(\lambda).$$

Comme dans le cas standard, on obtient :

$$\lambda = (T'_{szx})^{-1} (X - \hat{X}) + O\left(\|X - \hat{X}\|^2\right) \text{ avec}$$

$$z_k = \text{grad } F_k(0) \in \mathbb{R}^p \text{ et } T_{szx} = \sum_s d_k z_k x'_k.$$

Cette matrice est supposée de plein rang. Géométriquement, ceci signifie la chose suivante. Notons L_X le sous espace de \mathbb{R}^N engendré par les p variables coordonnées des x_k . L_Z est défini de la même façon à partir des z_k . La condition s'écrit géométriquement $L_X \cap L_Z^\perp = 0$.

Le cas particulier le plus simple est le cas linéaire où on prend simplement $F_k(\lambda) = 1 + z'_k \lambda$ avec un vecteur de variables "instrumentales" (ce choix de terme apparaîtra dans la suite) z_k .

On notera que la variance de l'estimateur calé $\hat{Y}_C = \sum_s d_k F_k(\lambda) y_k$ est, pour de gros échantillons, sensiblement la même que celle obtenue dans le cas linéaire. On a, dans ce cas :

$$\hat{Y}_C = \hat{Y} + (X - \hat{X})' T_{szx}^{-1} \sum_s d_k z_k y_k$$

$$= \hat{Y} + (X - \hat{X})' \tilde{\beta},$$

où $\tilde{\beta}$ est solution de :

$$\sum_s d_k z_k (y_k - x'_k \tilde{\beta}) = 0$$

On constate les faits suivants :

- $\tilde{\beta}$ est le vecteur des coefficients de la régression instrumentale (Fuller (1987) par exemple) utilisant les z_k comme instruments.

- Géométriquement, celle-ci s'interprète comme la projection, dans \mathbb{R}^N , du vecteur des y_k sur L_X le long de L_Z^\perp .

- Les poids de régression peuvent s'obtenir par minimisation d'une distance aux anciens poids.

- La variance de l'estimateur se calcule en utilisant les techniques des résidus. La différence avec le cas standard est qu'il faut utiliser les résidus de la régression instrumentale.

- L'estimateur de variance utilise le même principe.

- Les "instruments" z_k n'ont besoin d'être connus que sur l'échantillon : ils ne constituent pas une information auxiliaire.

2.3 Exemples

2.3.1 Estimateur par ratio.

X et x_k sont unidimensionnels. La variable instrumentale est la variable "gratuite" $z_k = 1$. L'équation de calage s'écrit :

$$X = \sum_s d_k x_k (1 + z_k \lambda) \text{ d'où } \tilde{\beta} = \frac{\hat{Y}}{\hat{X}} = \hat{R}$$

et les résidus valent $y_k - \hat{R} x_k$.

2.3.2 Estimateur par régression pondérée

Les instruments sont : $z_k = q_k x_k$.

2.3.3 Estimateur par régression optimal (Montanari (1987))

Les instruments sont :

$$z_k = \sum_{\ell \in s} \Delta_{k\ell} x_\ell \text{ avec } \Delta_{k\ell} = \frac{\pi_{k\ell}}{\pi_k \pi_\ell} - 1.$$

Dans le cas d'un plan stratifié avec sondage aléatoire simple dans chaque strate on a :

$$z_k = (x_k - \bar{X}_k) \frac{N_h^2}{n_h} \left(1 - \frac{n_h - 1}{N_h - 1} \right)$$

avec des notations habituelles.

2.3.4 Un exemple non-linéaire

Soit u_k une variable positive connue dans s , $\lambda' = (a, b, c)$ et $F_k(\lambda) = a + \exp(bu_k)u_k^c$.

On a bien $F_k(0) = 1$ et on trouve :

$$\frac{\partial F_k}{\partial a} = 1$$

$$\frac{\partial F_k}{\partial b} = u_k \exp(bu_k) u_k^c \Rightarrow \frac{\partial F_k}{\partial b}(0) = u_k$$

$$\frac{\partial F_k}{\partial c} = \text{Log } u_k \exp(bu_k) u_k^c \Rightarrow \frac{\partial F_k}{\partial c}(0) = \text{Log } u_k$$

3. NON-RÉPONSE POUR UNE ENQUÊTE EXHAUSTIVE : REpondÉRATION

3.1 Modèle de réponse et estimation

Le mécanisme de réponse est modélisé par un plan de sondage $q(r; \beta)$ où β est un paramètre inconnu de \mathbb{R}^p . Ce modèle nous fournit des poids d'extrapolation "à la Horvitz-Thompson", $\pi_k^{-1} = F_k(\beta)$, ainsi que des probabilités d'inclusion à l'ordre deux si nous en avons besoin pour exprimer la variance et l'estimer.

Le modèle le plus simple, et, à bien des égards, le plus naturel est le modèle de Poisson :

$$q(r; \beta) = \prod_{k \in r} F_k^{-1}(\beta) \prod_{k \in U-r} (1 - F_k^{-1}(\beta))$$

On peut aussi introduire des plans plus compliqués, avec par exemple des effets de grappe si on soupçonne qu'il y a des effets dus aux enquêteurs, par exemple.

La question maintenant est de savoir comment estimer β . La réponse, assez étonnante, est : peu importe ! Maximum de vraisemblance, méthodes des moments, Chi 2 - minimum, ce sera comme on voudra (ou presque !). Tout ce qui compte, c'est le fait qu'on arrive à un ensemble de p équations estimantes, qu'on résoudra quel que soit l'échantillon r possible ; de ce fait, ces équations vérifiées pour tout r constituent des statistiques dont la variance est nulle.

Exemple : Supposons qu'on estime le β du modèle de Poisson ci-dessus par la méthode du maximum de vraisemblance. On obtient les équations estimantes suivantes :

$$\sum_r F_k(\beta) z_k^* = \sum_U z_k^* \quad \text{avec}$$

$$z_k^* = \frac{\text{grad } F_k(\beta)}{F_k(\beta)(F_k(\beta) - 1)}$$

Si le modèle est spécifié plus précisément sous la forme d'un modèle linéaire généralisé :

$$F_k(\beta) = F(z_k' \beta) \text{ alors } z_k^* = z_k \frac{F'}{F(F-1)}$$

Si $F = 1 - \exp$ (modèle log-linéaire), on a simplement $z_k^* = z_k$.

De façon générale, il vaudra sans doute mieux se fier à un principe de calage et utiliser les équations estimantes sans biais suivantes :

$$\sum_r F_k(\beta) x_k = \sum_U x_k$$

ou avec un GLM: (*)

$$\sum_r F(z_k' \beta) x_k = \sum_U x_k$$

Il est immédiat que la solution de ces équations a une variance en $1/n$, n étant la taille de U . Mais on peut aussi en donner l'interprétation suivante :

$$X = \sum_r x_k F_k(\beta_0) G_k(\lambda)$$

où β_0 est la vraie valeur du paramètre,

$$\text{et, } G_k(\lambda) = \frac{F_k(\beta_0 + \lambda)}{F_k(\beta_0)}, G_k(0) = 1$$

Ces équations ne sont autres que des équations de calage qui appellent les commentaires suivants :

- $\beta = \beta_0 + \lambda$ est un estimateur de β_0 .

- Si, par hasard, on dispose d'une autre estimateur $\hat{\beta}_0$ de β_0 , on peut écrire $\beta = \beta_0 + (\hat{\beta}_0 - \beta_0) + \lambda$, et l'interprétation est la même car λ et l'erreur d'estimation $\hat{\beta}_0 - \beta_0$ sont de l'ordre de $1/n$.

- On n'a besoin de connaître les F_k que pour les répondants.

- L'effet sur la variance (et l'estimation de variance) est le même que celui obtenu dans le calage habituel.

3.2 Regardons plus en détail

D'après ce qu'on sait de la théorie généralisée du calage, la variance est celle de l'"estimateur linéaire" de la même famille, à savoir :

$\hat{Y} = \sum_r y_k F_k(\beta_0)(1+z_k'\lambda)$ avec λ vérifiant

$$X = \sum_r x_k F_k(\beta_0)(1+z_k'\lambda)$$

et $z_k = \text{grad } G_k(0) = \text{grad } \text{Log } F_k(\beta_0)$.

Si on utilise un modèle linéaire généralisé, on a $F_k(\beta) = F(z_k^* \beta)$ et

$$z_k = z_k^* \frac{\dot{F}(z_k^* \beta_0)}{F(z_k^* \beta_0)} = z_k^* (\text{Log } F(z_k^* \beta_0)) = z_k^* q_k.$$

On a alors : $\text{Var}(\hat{Y}) = \text{Var}(\sum_r \tilde{e}_k F_k(\beta_0))$ avec :

$$\tilde{e}_k = y_k - \tilde{B}' x_k, \text{ où } \tilde{B} \text{ est solution de :}$$

$$\sum_r z_k F_k(\beta_0) (y_k - \tilde{B}' x_k) = 0.$$

Dans le cas d'un GLM :

$$\sum_r q_k z_k^* F_k(\beta_0) (y_k - \tilde{B}' x_k) = 0.$$

Si $z_k^* = x_k$: $\sum_r q_k x_k F_k(\beta_0) (y_k - \tilde{B}' x_k) = 0$, et

l'interprétation est alors soit celle des variables instrumentales, soit celle des moindres carrés pondérés.

3.3 Des exemples

3.3.1 Redressement par ratio

x_k est une variable positive, et $x_k = 1$. Autrement dit les réponses manquent au hasard mais je cale sur le total X des x_k . Alors : $\hat{Y} = X \frac{Y_r}{X_r}$.

Avec le modèle Poisson on obtient la variance suivante : $\frac{X}{X_r} (\frac{X}{X_r} - 1) \sum_r (y_k - R x_k)^2$.

Si $x_k = 1$ on obtient : $\frac{N}{n} (\frac{N}{n} - 1) \sum_r (y_k - \bar{y})^2$

3.3.2 Poststratification et raking-ratio formel.

Tout va de soi.

3.3.3 Un exemple plus original

Le vecteur x_k est l'indicateur d'une variable qualitative à I modalités $i = 1$ à I . L'effectif N_i est supposé connu dans U . z_k est un vecteur indicateur d'une autre variable qualitative de même dimension indicé par a .

On peut imaginer, par exemple, que x_k est une situation connue dans la base de sondage et z_k la situation mesurée lors de l'enquête (exhaustive, rappelons le, mais entachée de non réponse). On note R_{ia} l'effectif des répondants classés à la modalité i de x_k et a de z_k . Le vecteur $\beta = (\dots, \beta_a, \dots)$ des paramètres du modèle de réponse vérifie les équations suivantes : $N_i = \sum_a R_{ia} (1 + \beta_a)$, ce qui permet

d'estimer β . La variance de l'estimateur calé vaut : $\sum_a \beta_a (1 + \beta_a) \sum_i \sum_{k \in r_{ia}} (y_k - \hat{Y}_i)^2$ où \hat{Y}_i est l'estimation de la moyenne des y_k pour k classé dans la modalité i de x_k .

4. NON RÉPONSE APRÈS ÉCHANTILLONNAGE

4.1 Position du problème

Un plan de sondage p sur la population U fournit un échantillon s auquel sont associés les poids d'extrapolation d_k . Conditionnellement à s , la non réponse est régie par un plan de sondage $q(r|s; \beta)$ fournissant l'échantillon r . Les poids d'extrapolation de r vers s sont de la forme $F_k(\beta_0)$ conformément à ce qui se passait dans le paragraphe 3. Si les x_k sont connus sur s , l'estimation de β conformément au principe de calage doit respecter l'information disponible sur s , soit $\hat{X}_s = \sum_s d_k x_k$. Ceci nous

conduit donc aux équations estimantes :

$$\hat{X}_s = \sum_s d_k x_k = \sum_r d_k x_k F_k(\beta_0).$$

Comme cela est bien connu (Särndal, Wretman (1987)), nous avons affaire à un sondage en deux phases. La variance comporte deux termes qui s'estiment séparément. Dans la suite nous supposons que nous disposons d'un logiciel (comme POULPE, Caron, (1998), Petit (1998)) qui calcule de façon automatique ces deux formes quadratiques :

$$\hat{\text{Var}}(\hat{Y}) = Q_1(y_r) + Q_2(y_r), \text{ où } y_r = \{y_k ; k \in r\}.$$

Si nous avons estimé β_0 à l'aide de l'équation (4.1.1) ci-dessus, la variance sera (estimée par) :

$$\hat{\text{Var}}(\hat{Y}_1) = Q_1(y_r) + Q_r(e_r^{zx})$$

où e_r^{zx} est l'ensemble des résidus de la régression instrumentale effectuée dans s et d'équations

normales:

$$\sum_s z_k (d_k y_k - d_k x_k' B) = 0 \quad (z_k = \frac{\text{grad } F_k(\beta_0)}{F_k(\beta_0)})$$

Ces équations estiment la régression d'équations normales dans U : $\sum_U z_k (y_k - x_k' B) = 0$.

L'estimation dans r est donnée par :

$$\sum_r d_k F_k(b) z_k (y_k - x_k' B) = 0$$

résidus

4.2 Cas où X est connu

Dans ce cas, on peut caler l'estimateur précédent sur cette information. La variance (estimée) prend alors la forme : $\hat{\text{Var}}(\hat{Y}_1) = Q_1(e_r^{xx}) + Q_2(e_r^{zx})$. Si on regarde le cas $z_k = x_k$, les deux régressions deviennent identiques et les résidus utilisés dans Q_1 et Q_2 sont les mêmes.

4.3 Cas où X est connu mais pas \hat{X}_s

C'est un cas fréquent. Le total X est connu de façon externe ; cependant \hat{X}_s ne peut être calculé à cause de la non réponse. Un cas typique et fréquent est celui de la poststratification formelle où on utilise l'estimateur aux formes équivalentes suivantes :

$$\hat{Y}_{\text{postform}} = \sum_h N_h \frac{\sum_{th} d_k y_k}{\sum_{th} d_k} = \sum_h \frac{N_h}{\hat{N}_h} \frac{\hat{N}_h}{\sum_{th} d_k} \sum_{th} d_k y_k$$

Dans le cas général, examinons ce qui se passe si nous résolvons les équations : $X = \sum_r d_k x_k F_k(\beta)$.

En fait, les choses se passent exactement comme dans la partie 3. Il suffit de réécrire cela sous la forme :

$$X = \sum_r d_k F_k(\beta_0) x_k G_k(\lambda)$$

En appliquant les résultats établis à ce moment, on obtient donc que la variance (estimée) vaut :

$$\hat{\text{Var}}(\hat{Y}_3) = Q_1(e_r^{zx}) + Q_2(e_r^{zx})$$

Si $F_k(\beta) = F(z_k^* \beta)$ alors $z_k = z_k^* \frac{\dot{F}(z_k^*)}{F(z_k^*)} = z_k^* q_k$.

Si $F_k(\beta) = F(x_k' \beta)$ on obtient l'estimateur calé "standard" avec des poids de régression q_k . Si, en plus, $F = \text{exp}$, alors on a tout bonnement $q_k = 1$.

5. DOUBLE CALAGE POUR CORRIGER LA NON RÉPONSE

C'est le final avec toute la troupe !

5.1 Le problème

On dispose de l'information auxiliaire suivante :

$X_o = \sum_U x_{ok}$, $\dim X_o = p_o$ est donc connu au niveau de la population U.

$\hat{X}_1 = \sum_s d_k x_{1k}$, $\dim X_1 = p_1$ est un estimateur utilisable au niveau de l'échantillon s (les variables de x_{1k} ne sont pas affectées par la non réponse). On connaît aussi, pour $k \in r$, les "fonctions de calage et non réponse associée" $F_k(\beta)$ ou β est un paramètre de R^p , $p = p_o + p_1$. Sans chercher à trop tourner autour du pot, compte tenu de ce qui a été vu jusqu'ici, nous allons directement nous demander ce qui se passe si nous résolvons les équations de calage :

$$\begin{pmatrix} X_o \\ \hat{X}_1 \end{pmatrix} = \sum_r d_k \begin{pmatrix} x_{ok} \\ x_{1k} \end{pmatrix} F_k(\beta) = \sum_r d_k x_k F_k(\beta_0) G_k(\lambda)$$

avec $G_k(\lambda) \cong 1 + z_k' \lambda$ et $z_k = \frac{\text{grad } F_k(\beta_0)}{F_k(\beta_0)}$

Encore une fois, $\beta = \beta_0 + \lambda$ peut s'interpréter comme un estimateur de β_0 intégrant le terme de calage en plus de l'erreur d'estimation.

5.2 Le résultat

Les équations estimantes étant sans biais sous le modèle de réponse, l'estimateur repondéré est évidemment sans biais "asymptotiquement".

Le résultat le plus intéressant concerne la variance d'un tel estimateur. L'information utilisée au cours des deux phases d'estimation n'étant pas la même, il est naturel d'obtenir un résultat qui rend compte de ce fait.

RÉSULTAT :

L'estimateur \hat{Y}_w utilisant les poids déduits de l'équation (5.1.1) admet pour variance (estimée) :

$$\hat{\text{Var}}(\hat{Y}_w) = \sum_r d_k F_k(\beta) y_k = Q_1(e_r^s) + Q_2(\tilde{e}_r)$$

où :

$$\tilde{\epsilon}_k = y_k - x_{ok} \hat{\beta}_0 - x_{1k} \hat{\beta}_1 \quad (5.2.1)$$

vérifie les équations normales instrumentales:

$$\sum_r d_k F_k(\beta) z_k (y_k - (x_{ok} \quad x_{1k}) \begin{pmatrix} \tilde{B}_0 \\ \tilde{B}_1 \end{pmatrix}) = 0$$

(équation estimante pour

$$\sum_U z_k (y_k - (x_{ok} \quad x_{1k}) \begin{pmatrix} \tilde{B}_0 \\ \tilde{B}_1 \end{pmatrix}) = 0)$$

et $e_r^s = y_k - x_{ok} B_0^s$ vérifie les équations instrumentales normales :

$$\sum_r d_k F_k(\beta) x_{ok}^* (y_k - x_{ok} B_0^s) = 0 \quad (5.2.2)$$

avec les instruments : $x_{ok}^* = T_{rx_0z} T_{zz}^{-1} z_k$

$$(\text{avec } T_{ruv} = \sum_r d_k F_k(\beta) u_k v_k')$$

Un cas particulier assez général :

Il peut arriver assez souvent que x_{ok} résulte d'une transformation linéaire de z_k . Dans ce cas B_0^s n'est autre que le vecteur des coefficients de la régression des moindres carrés. On obtient donc aussi que $x_{ok}^* = x_{ok}$.

Un cas très particulier très général :

Si $z_k = \begin{pmatrix} x_{ok} \\ x_{1k} \end{pmatrix}$, tout redevient assez simple. On a

$$\begin{pmatrix} \tilde{B}_0 \\ \tilde{B}_1 \end{pmatrix} = \begin{pmatrix} B_0 \\ B_1 \end{pmatrix} \text{ car il s'agit des moindres carrés. On}$$

trouvera aussi que $B_0^s = B_{0o}$, estimateur des moindres carrés de la régression de y sur x_o . Ce dernier résultat est une légère amélioration de celui de F. Dupont (1995).

5.3 Une ébauche de preuve

La formule (5.2.1) donnant la partie conditionnelle à s de la variance est assez naturelle et résulte directement de ce qui a été établi dans la partie 3.

Par ce qui est de (5.2.2), le cheminement et le suivant. On remarque que la régression instrumentale de y sur (x_o, x_1) a les mêmes coefficients que la régression des moindres carrés sur (x_o^*, x_1^*) qui sont les projections de x_o et x_1 sur L_Z le long de L_Z^\perp . On exprime ensuite la régression des moindres carrés des x_1^* sur les x_o^* .

On constate alors que $E(\hat{Y}_w | s)$ prend la forme de l'estimateur par régression instrumentale dont les équations normales sont (5.2.2). La formule de variance en résulte.

6. QUELQUES IDÉES SUR L'IMPUTATION DES VALEURS MANQUANTES

La façon habituelle de concevoir les questions d'imputation pour valeurs manquantes y_k consiste à estimer par un modèle la loi de probabilité que y_k est censée suivre conditionnellement à un cofacteur x_k . Ce modèle est estimé à partir des données où x_k et y_k sont simultanément présents. Deux méthodes s'opposent alors. L'une consiste à imputer l'espérance de y_k . Elle fournit le meilleur prédicteur de $\sum_U y_k$ ou $\sum_s w_k y_k$.

Elle a l'inconvénient de mal restituer la dispersion des y_k et donc de fausser des statistiques liées à la fonction de répartition des y_k . L'alternative consiste à imputer une valeur de y_k dans la loi estimée. L'inconvénient est d'ajouter une variance arbitraire et souvent importante (voir par exemple Oh et Scheuren (1983)) à l'estimation du total des y_k . Lorsque ces méthodes sont basées sur des modèles non paramétriques elles se ramènent souvent à une imputation par donneur choisi au hasard avec un certain jeu de probabilités. Formellement, c'est toujours le cas lorsque y_k est une variable qualitative représentée par le vecteur de ses indicatrices.

Une solution qui possède les avantages des deux méthodes consiste alors à imputer par échantillonnage dans les répondants en respectant le meilleur total estimé. C'est une variante d'échantillonnage équilibré.

L'exemple le plus simple est le suivant. y_k est une variable binaire pouvant valoir 0 ou 1. Le modèle nous livre des probabilités P_k pour que y_k vaille 1. Si l'échantillon est à probabilités égales (ou si l'enquête est exhaustive comme au paragraphe 2) la meilleure prédiction est basée sur $\sum_o P_k = n$ qu'on supposera

entier. La question est alors d'échantillonner n fois la valeur 1 et $|Q| - n$ fois la valeur 0. C'est un échantillonnage à probabilités inégales de taille fixée. Cette situation se généralise à de nombreux cas et introduit une connection très riche entre l'imputation et un échantillonnage équilibré de donneurs.

Une autre approche mérite attention.

Formellement, on peut toujours repondérer pour compenser à des données manquantes concernant une variable particulière. L'imputation est un artifice destiné à pallier aux complications qui naîtraient de l'usage de plusieurs familles de poids différentes associées aux différentes variables d'intérêt. L'astuce suivante permet de rester fidèle à cette vision de repondération tout en gardant la commodité de l'imputation.

La correction pour non-réponse fait apparaître des poids modifiés de la forme $w_k = g_k d_k$. Les corrections de poids g_k s'interprètent comme des inverses de probabilité de réponse. Ce sont des quantités généralement supérieures à 1, rarement supérieures à 2.

L'estimation des g_k est fondée sur un modèle de mécanisme de réponse qui conduit à un ensemble (E) d'équations estimantes. Quitte à ajouter un paramètre au modèle initial, on peut toujours faire en sorte que (E) contienne une équation de normalisation de la forme $\sum_r g_k = n$.

Une piscine de donneurs peut alors être constitué comme suit (on peut supposer que tous les g_k sont supérieurs à 1):

- Elle comporte $[g_k] - 1$ copies de l'unité k ($[g_k]$ est la partie entière de g_k).

- $\sum_0 g_k - [g_k]$ est un entier p . On échantillonne p unités k avec des probabilités proportionnelles à $g_k - [g_k]$ en respectant des contraintes supplémentaires liées à (E).

La piscine contient autant d'unités (duplicata compris) que d'unités à imputer et l'attribution des valeurs de la piscine peut se faire au hasard parmi les unités présentant une non-réponse.

Il est facile de voir que le jeu de valeurs ainsi complété vérifie exactement l'ensemble des équations (E) et possède donc les mêmes propriétés statistiques (réduction de variance comprise) que de jeux de données qu'on aurait obtenu par repondération.

REMERCIEMENTS

Je tiens à remercier la Société de Statistique du Canada pour son aimable invitation et pour l'accueil chaleureux dont j'ai été l'objet.

RÉFÉRENCES

DEVILLE, J.C, et SÄRNDAL, C.E.: Calibration estimators in survey sampling, Journal of the American Statistical Association, Vol 87, pp 376-382 (1992)

DEVILLE, J.C, SARNDAL, C.E, et SAUTORY, O.: Generalized raking procedures in survey sampling, Journal of the American Statistical Association, Vol 88, pp 1013-1020 (1993)

FULLER, W.A., Measurement Error Models, Wiley (1987)

CARON, N.: Le logiciel POULPE: aspects méthodologiques, INSEE: Actes des Journées de Méthodologie (1998)

PETIT, J.N.: Le logiciel POULPE: modélisation informatique, INSEE: Actes des Journées de Méthodologie (1998)

DUPONT, F.: Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire, Techniques d'enquête, vol. 21, n° 2, pp.141-150 (1995).

MONTANARI, G.E.: Post sampling efficient prediction in large scale surveys, International statistical review, Vol 55, pp191-202 (1987)

SÄRNDAL, C.E. et SWENSSON, B.: A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse, International statistical review, Vol 55, pp 279-294 (1987)

DEVILLE, J.C, et DUPONT, F. : Non-réponse : Principes et Méthodes, INSEE : Actes des Journées de Méthodologie Statistique de 1993, (1996)

OH, H.L et SCHEUREN, F.J : Weighting Adjustments for Unit Nonresponse, dans Incomplete Data in Sample Surveys, Vol 2, pp143-184 (1983)