IMPUTATION PAR PRÉDICTION OU IMPUTATION AVEC ALÉA ?

Jean-Claude DEVILLE

ENSAI/CREST
Laboratoire de Statistique d'Enquête, Campus de Ker-Lan

1. Rappel sur la correction de la non-réponse par repondération

Le but est d'obtenir de nouveaux poids $w_k=d_k * g_k$ plus 'représentatifs', c'est à dire où les g_k sont des estimations des inverses de probabilités de réponse. On tient compte alors de ce que la non-réponse peut être vue comme une phase supplémentaire d'échantillonnage, mais non contrôlée, et donc obéissant à des paramètres qu'on se doit d'estimer dans le cadre d'un modèle de réponse.

Pour toute variable dérivée de y , tant linéaire (comme $y_k * \mathbf{1}_k^D$, où $\mathbf{1}_k^D$ est l'indicatrice d'un domaine D) que non-linéaire (comme un fractile, $\mathbf{1}(y_k < t)$), où, plus généralement une fonctionnelle construite à partir de la fonction de répartition, cette méthode permet de trouver des estimateurs à biais négligeable, *si le modèle de réponse est juste et formalise correctement le* ***mécanisme de réponse.*** De plus**,** les liaisons entre variables sont respectées et on peut donc, sans restriction ni arrière pensée, utiliser les données repondérées pour fabriquer des tableaux croisés, estimer des corrélations ou se livrer à des analyses économétriques (qui sont, de fait, basées sur l'estimation de corrélations).

Bref, la correction par repondération de la non-réponse offre une solution presque parfaite aux problèmes statistiques que peut poser l'analyse des données d'enquête.

Malheureusement cette technique ne peut pas s'appliquer quand la non-réponse affecte de façon différente les diverses variables utiles. On recourt alors, classiquement, à la technique d'imputation qui consiste à remplacer les valeurs manquantes par des valeurs plausibles. On obtient ainsi des données à structure rectangulaire qui se prêtent à toutes les opérations possibles d'analyse des données de l'enquête. Il se trouve que se posent alors de nombreux problèmes cachés et que la fiabilité des résultats demande une analyse approfondie de la nature des transformations qu'on a fait subir aux données. On examinera ici des choses de bases : génère-t-on des biais ? quelle est la variance des données produites et comment l'estime-t-on ?

On verra que la réponse est assez différente selon que l'on impute une prédiction de la valeur manquante où une valeur aléatoire dans une loi de probabilité qu'elle est censée suivre.

2. Principe de la correction par imputation

Le point de vue est dans cette approche entièrement différent. Les y_k manquant sont considérés comme des variables aléatoires (dans un modèle dit de superpopulation) dont on estime la loi $£_k$ à partir de l'échantillon *r* de répondants (de façon paramétrique ou non-paramétrique). L'estimation du modèle ne dépend donc que des valeurs de la variable d'intérêt sur l'ensemble des répondants (et, en général, de la valeur de variables auxiliaires sur l'ensemble de l'échantillon).

On a alors le choix entre:

-Imputer la meilleure prédiction de y_k , c'est à dire l'espérance $\hat{y}_k$ dans la loi estimée de $£_k$ (attention, pour une variable 0-1, cette espérance est une probabilité!).

-Imputer un aléatoire $\tilde{y}_k$ tiré aléatoirement dans cette loi (par exemple 0 ou 1 dans le cas d'une variable 0-1).

L'estimateur est $\hat{Y}^* = \sum_s d_k \; y_k^*$ où $y_k^* = y_k$ dans *r*, et $\hat{y}_k$ ou $\tilde{y}_k$ dans *o*, selon la méthode choisie.

- *Dans le premier cas, les valeurs imputées dépendent de r de façon parfaitement déterministe.*

- *Dans le second, la loi estimée dépend de r de façon déterministe. Par contre les valeurs imputées dépendent d'un nouveau mécanisme aléatoire indépendant des aléas d'échantillonnage (mécanisme de réponse compris). Sous ce mécanisme (formellement, conditionnellement à r) l'espérance de* $\tilde{y}_k$ *vaut* $\hat{y}_k$ *(qui ne dépend que de r), mais une variance supplémentaire 'parasite' affecte l'estimateur.*

3. Exemple de base :

L'échantillonnage est un sondage aléatoire simple (SAS) de taille *n* et le mécanisme de réponse est Bernoullien conduisant à un échantillon de répondants de taille *m.* On ne dispose d'aucune variable auxiliaire (à l'exception de la constante, variable gratuite). Les y_k défaillant ont donc tous la même loi qu'on estime non-paramétriquement. La fonction de répartition estimée est donc celle des répondants. L'espérance de la loi estimée est $\bar{y}_r$, moyenne des *y* parmi les répondants, la variance corrigée de y_k est donc $s_r^2 = \dfrac{\sum_r (y_k - \bar{y}_r)^2}{m-1}$.

Dans le premier cas (imputation déterministe) on aura donc : $\hat{y}_k = \bar{y}_r$, moyenne des *y* parmi les répondants. La moyenne des données imputées sera toujours $\bar{y}_r$. Il est alors facile de vérifier que la moyenne générale est encore estimée par $\bar{y}_r$, moyenne des *y* parmi les répondants et que la moyenne sur un domaine est estimée sans biais (sous le modèle de réponse). En revanche la médiane des *y* ne pourra pas être estimée décemment!

Dans le second cas (imputation avec aléa), les $\tilde{y}_k$ seront tirés au hasard (avec probabilités égales) parmi les y_k c'est à dire qu'on imputera par *hotte-dekke.* Chaque y_k aura une espérance égale à $\bar{y}_r$ et, ***conditionnellement à r,*** une variance d' « échantillonnage parasitaire » égale à s^2_r. La moyenne des données imputées aura encore une espérance égale à $\bar{y}_r$, mais additionnée d'une erreur aléatoire parasite qui dépend de la loi jointe (dans le tirage pour imputation) des $\tilde{y}_k$. Dans la pratique, on utilise souvent des tirages indépendants, ce qui maximise cette variance superflue !

Par contre, contrairement au premier cas, la médiane (par exemple) sera estimée proprement, c'est à dire avec un biais négligeable en *1/m.*

Quand la variable est de type 0-1, le prédicteur n'est autre que la fréquence *f* des valeurs 1 parmi les répondants. Comme ce n'est pas une valeur possible, on comprend que la préférence aille souvent dans ce cas à l'imputation aléatoire (bien que celle ci ne soit fondée que sur un argument « esthétique »).

L'imputation aléatoire consiste à imputer des 1 avec la probabilité f. Ceci posé, on aimerait bien, quand même, que la proportion de 1 soit de *f* parmi les valeurs imputées. Autrement dit, au lieu d'imputer des 1 indépendamment (échantillon bernoullien), on réalisera un sondage aléatoire simple (de taille fixe *f(n-m)*, donc) pour déterminer les unités imputées à 1. Autrement dit encore, les donneurs de '1' seront déterminés par cette variante élémentaire d'échantillonnage équilibré qu'est l'échantillonnage de taille fixe.

4. Que peut-on estimer (presque) sans biais avec des données imputées ?

On supposera, bien entendu, que les modèles utilisés rendent parfaitement compte de la réalité des données.

Dans le cas de l'imputation avec des prédicteurs, il est clair que, si les prédicteurs sont sans biais, l'estimation du total des *y* sera également sans biais. On pourrait s'attendre à ce que ce soit encore vrai pour toute combinaison linéaire de la forme $\sum_U c_{k,l}\ y_l$. Ce serait le cas si les prédicteurs étaient disponibles pour toute la population *U*. C'est la technique des estimateurs par prédiction, bien étudiée par ailleurs, qui considère comme une non-réponse toute unité de *U-r*, qu'elle ait été échantillonnée ou pas dans *s*. Malheureusement, en général, la prédiction s'appuie sur des variables auxiliaires présentes dans *s* de sorte qu'on ne peut utiliser que l'estimateur $\hat{Y}^* = \sum_s d_k\ y_k^*$ décrit ci-dessus. Il est facile de voir que les seules statistiques de la forme $\sum_U c_{k,l}\ y_l$ qu'on puisse estimer sont celles où la matrice des c_{kl} est diagonale. Autrement dit, des quantités de la forme $\sum_U c_k\ y_k$ seront estimée (presque) sans biais par $\sum_s d_k c_k\ y_k^*$. C'est ce que nous appellerons dans la suite une transformation linéaire de y. Apparemment un peu limité, ce cas recouvre néanmoins des applications importantes comme les totaux sur domaine, où on a $c_k = 1(k \in D)$ pour un domaine *D* de *U*.

En revanche, comme on va le voir tout de suite, l'estimation de transformées non-linéaires de *y* est désespérée. Dans le cas de l'imputation avec des aléas, la situation l'est moins.

Limitons nous à l'estimation de la fonction de répartition de $y : F_y(t) = \sum_U 1(y_k < t)$. Cela suffit, en effet, pour traiter toute fonctionnelle de la variable *y*, en vertu du principe de substitution. Sans finasser, on examinera donc les propriétés de $\hat{F}_y^*(t) = \sum_s d_k 1(y_k^* < t)$. Si le modèle est juste, quand on impute avec aléa, chacune des variables $1(y_k^* < t)$ suit la loi induite par $\pounds_k$, que *k* soit dans *r* ou dans *o* (ce n'est plus vrai quand on impute l'espérance selon $\pounds_k$!). Si le modèle est juste, donc, $\hat{F}_y^*(t)$ sera un estimateur sans biais de la fonction de répartition des *y*.

5. Système d'imputation paramétrique

La loi $\pounds_k$ est supposée paramétrique. Autrement dit on supposera que la loi de y_k pour une variable auxiliaire « explicative » x_k donnée dépend d'un paramètre β à p dimensions réelles β. Ce paramètre sera estimé a partir des données de r par un système de p équations estimantes de la forme:

$$\sum_r u_k(y_k, x_k; \beta) = 0 \qquad (1)$$

où les u sont donc des fonctions à valeurs dans $\mathbf{R}^p$ et x_k une information présente dans s.

Une des coordonnées des u_k pourra être, par exemple,, l'équation estimante exprimant que la somme des prédicteurs sur r doit égaler la somme des valeurs observées. Si le modèle de réponse n'est pas un sondage aléatoire simple et que des probabilités de réponse P_k ont été estimées, on pourra aussi utiliser des équations de la forme:

$$\sum_r P_k^{-1} u_k(y_k, x_k; \beta) - \sum_s v_k(x_k; \beta) = 0 \quad \text{où } v_k(x_k; \beta) = u_k(\hat{y}_k(x_k; \beta), x_k; \beta) \qquad (2)$$

de façon à obtenir le même résultat pour un virtuel estimateur repondéré que pour l'estimateur imputé. Par exemple, si on désire ne pas modifier l'estimateur du total des y obtenu par repondération on utilisera l'équation estimante $\sum_r \frac{d_k}{P_k} y_k - \sum_s d_k \hat{y}_k(x_k; \beta) = 0$.

Exemple : on utilise un unique paramètre R et une seule équation estimante : $\sum_r d_k (y_k - R x_k) = 0$. Le résultat est la classique imputation par ratio, et, si le modèle de réponse est un SAS, l'estimateur correspondant est l'estimateur par ratio.

Remarque : En dépit des apparences, les équations (2) sont bien un cas particulier des équations (1). Il suffit de poser $u_k^* = P_k^{-1} u_k(y_k, x_k; \beta) - \frac{m}{n} \sum_s v_k(x_k; \beta)$ pour s'y ramener.

6. Estimation de la variance du paramètre d'ajustement

Le choix des équations estimantes est important pour la variance de l'estimateur obtenu. D'autre part, si les P_k ont été estimées, dans tous les cas la variance 'design based ' du paramètre d'ajustement β est calculable et estimable. Partons, en effet, des équations (1) et notons $\hat{\beta}_r$ sa solution (supposée bien définie, unique et régulière) pour l'échantillon de répondants r. L'espérance de $\sum_r u_k(y_k, x_k; \beta)$ pour le modèle du mécanisme de réponse vaut $\sum_s P_k u_k(y_k, x_k; \beta)$. Notons β_0 la solution de $\sum_s P_k u_k(y_k, x_k; \beta_0) = 0$. On obtient par linéarisation au voisinage de β_0 :

$$\sum_r u_k(y_k, x_k; \beta_0) + \sum_s P_k \frac{\partial}{\partial \beta} u_k(y_k, x_k; \beta_0)(\hat{\beta}_r - \beta_0) = 0$$

de sorte que la variable linéarisée de $\hat{\beta}_r$ est :

$$lin(\hat{\beta}_r) = \left(\sum_s P_k \frac{\partial}{\partial \beta} u_k(y_k, x_k; \beta_0) \right)^{-1} u_k(y_k, x_k; \beta_0)$$

On obtient donc la variance conditionnelle à *s* de $\hat{b}_r$ en portant cette expression dans celle de la variance du modèle de réponse (en général un modèle poisonnien ou de sondage aléatoire simple) et on l'estime en portant dans l'estimateur de cette variance conditionnelle l'approximation habituelle $\left(\sum_r \frac{\partial}{\partial b} u_k(y_k, x_k; \hat{b}_r) \right)^{-1} u_k(y_k, x_k; \hat{b}_r)$.

7. Variance conditionnelle de l'estimateur par prédiction : comment ça marche ?

Avec les notations déjà introduites, l'estimateur avec valeurs imputées par prédiction, dans le cas paramétrique, va s'écrire :

$$\hat{Y}_{imp,pred} = \sum_r d_k y_k + \sum_o d_k \hat{y}_k(x_k; \hat{b}_r)$$

Sa variance (au sens design based) est en principe assez facile à calculer.

En effet, soient $e_k = y_k - \hat{y}_k(x_k; b_0)$ les 'résidus vrais' et $\tilde{e}_k = y_k - \hat{y}_k(x_k; \hat{b}_r)$ les résidus empiriques. On a aussi :

$$\hat{Y}_{imp,pred} = \sum_r d_k \tilde{e}_k + \sum_s d_k \hat{y}_k(x_k; \hat{b}_r)$$

Par linéarisation au voisinage de b_0 on constate que la variance de l'estimateur n'est autre que celle de :

$$\sum_r d_k e_k + \sum_s d_k \hat{y}_k(x_k; b_0) + (\sum_s d_k l_k)(\hat{b}_r - b_0)$$

avec $l_k = \frac{\partial}{\partial b} \hat{y}_k(x_k; b_0)$ (noté comme un vecteur ligne).

La variance conditionnelle du terme central est nulle d'où on déduit que la variable linéarisée (pour la variance conditionnelle à *s*) de l'estimateur vaut : $e_k + (\sum_s d_k l_k)\, lin(\hat{b}_r)$.

On en déduit de la façon habituelle la variable approximative permettant le calcul effectif de l'estimateur de la variance conditionnelle.

Pour obtenir la variance totale il faut ajouter l'estimation de $Var(\sum_s d_k \hat{y}_k(x_k; \hat{b}_r))$, ce qui ne pose pas de problème supplémentaire.

Exemple : Examinons le cas de l'imputation par ratio.

8. L'estimation avec aléa : comment ça marche ?

Les valeurs imputées sont de la forme $\hat{y}_k + e_k^*$ où les e_k^* sont des résidus d'espérance nulle dans la loi estimée. L'estimateur imputé peut donc s'écrire :

$$\hat{Y}_{imp,alea} = \sum_r d_k y_k + \sum_o d_k \, \hat{y}_k(x_k; \hat{b}) + \sum_o d_k e_k^*$$

Il est (presque) sans biais et sa variance est facile à comprendre, car:

$$Var(\hat{Y}_{imp,alea}) = Var(\hat{Y}_{imp,pred}) + Var_{imp}(\sum_o d_k e_k^*)$$

où le second terme est celui qui résulte du processus (aléatoire) d'imputation.

On voit donc clairement apparaître que l'estimateur imputé aléatoirement a pour espérance un estimateur imputé par prédiction mais qu'il est doté d'une variance supplémentaire due au mécanisme aléatoire de l'imputation. Cet accroissement artificiel peut être réduit si on tire les résidus e_k^* dans une loi jointe ayant les $£_k$ pour marginales *et présentant des corrélations négatives,* typiquement par des échantillonnages équilibrés (le retour de cube !).

Cependant, à ce prix, souvent élevé, d'une variance accrue, l'imputation avec aléa permet une estimation (presque) sans biais du total d'une transformée quelconque de la variable *y*.

9. Conclusions

Dans tous les cas, et la mise au point de l'estimateur imputé, et, surtout, l'évaluation et l'estimation de la variance ne peuvent que difficilement faire l'économie d'une étude soigneuse suivie d'une modélisation du mécanisme de réponse.

L'imputation par prédiction ne demande que l'estimation d'un prédicteur des valeurs manquantes. En revanche, l'estimation avec aléa repose sur l'estimation de *la loi* $£_k$ de chacune de ces valeurs. Cette procédure est donc moins robuste (et beaucoup plus risquée!) que la première.

L'usage de données imputées est assez utile comme substitut de la repondération pour produire des statistiques simples (totaux ou fonction de totaux) de la variable imputée. L'imputation avec des prédicteurs permet aussi l'estimation sans biais de transformations linéaires de la variable imputée (en pratique des totaux restreints à un domaine). L'imputation avec des aléas autorise même des transformations non linéaires de cette variable (la fonction de répartition peut être estimée sans biais substantiel). Tout cela, bien sur, sous l'hypothèse que les données sont générées conformément au modèle de superpopulation qui sert de support à l'imputation.

En revanche, les corrélations entre variables sont altérées par les formes d'imputation (de loin les plus habituelles) que nous avons étudiées. Il est donc très dangereux de les utiliser pour l'élaboration de statistiques croisant deux ou plusieurs variables. Il est, en particulier, fondamentalement pervers d'utiliser des données imputées dans toute analyse de nature économétrique : on s'est donné l'illusion d'un enrichissement des données, alors qu'en réalité, on les appauvrissait.