Imputation par prédiction ou imputation avec aléa ?

Jean-Claude Deville
ENSAI/CREST
Laboratoire de Statistique d'Enquête
Campus de Ker-Lann-35170-BRUZ

Correction par repondération

But: obtenir de nouveaux poids $w_k = d_k * g_k$ les g_k étant des estimations d'inverses de probabilités 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 du domaine D) que non-linéaire (comme un fractile, $\mathbf{1}(y_k < a)$), 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.***

Correction par imputation

Les y_k manquant sont considérés comme des variables aléatoires dont on estime la loi $£_k$ à partir de l'échantillon *r* de répondants (de façon paramétrique ou non-paramétrique). On a alors le choix entre:

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

-Imputer un aléatoire $y^{\sim}_k$ dans cette loi.

-L'estimateur est $\sum_s d_k y^*_k$ -où $y^*_k = y_k$ dans *r* et $y^{\wedge}_k$ ou $y^{\sim}_k$ dans *o*.

-Dans le premier cas, les valeurs imputées ne dépendent que de r

-Dans le second, leur espérance dépend de r, mais une variance parasite s'ajoute.

Exemple de base:

Aucune variable auxiliaire, mécanisme de réponse Bernoullien sur un sondage aléatoire simple. Les y_k défaillant ont donc tous la même loi qu'on estime non-paramétriquement: sa fonction de répartition est celle des répondants.

Donc, dans le premier cas on aura: $\hat{y}_k = \bar{y}_r$, moyenne des répondants. La moyenne des données imputées sera toujours $\bar{y}_r$, la moyenne d'un domaine sera estimée sans biais (sous le modèle de réponse). Par contre la médiane ne pourra pas être estimée décemment!

Dans le second, $\tilde{y}_k$ sera tiré au hasard (avec probabilités égales) parmi les y_k c'est à dire qu'on imputera par *hotte-dekke*. La moyenne des données imputées aura une espérance égale à $\bar{y}_r$ et, ***conditionnellement à r,*** une variance d'échantillonnage parasitaire. Par contre, la médiane (par exemple) sera estimée proprement (biais négligeable).

Exemple de base (suite):

Quand la variable est de type 0-1, le prédicteur n'est autre que la fréquence *f* des valeurs 1 parmi les répondants. Ce n'est pas une valeur possible!

L'imputation aléatoire consiste à imputer des 1 avec la probabilité *f* .

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, donc) pour déterminer les unités imputées à 1. Autrement dit, encore, les donneurs de '1' seront déterminés par une variante élémentaire d'échantillonnage équilibré.

Système d'imputation paramétrique

Un paramètre *p*-dimensionnel β indice la loi des y_k pour x_k donné.

Il est estimé a partir des données de *r* par un système de *p* équations estimantes de la forme:

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

où les *u* sont des fonctions à valeurs dans $\boldsymbol{R}^p$ et x_k une information présente dans s. Une des coordonnées pourra être $y_k - \hat{y}_k(x_k;\beta)$ pour la loi $£_k$, exprimant que la somme des prédicteurs sur *r* doit égaler la somme des valeurs observées.

Système d'imputation paramétrique (suite)

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:

$$\sum_r u_k(y_{k,}\ x_k) - \sum_s v_k(x_k;\beta) = 0$$

de façon à obtenir le même résultat pour un virtuel estimateur repondéré que pour l'estimateur imputé. Le choix des équations estimantes est important pour la variance de l'estimateur obtenu.

Exemple:

$$\sum_r \frac{d_k}{P_k} y_k - \sum_s d_l \hat{y}_k(x_k;\beta) = 0$$

Système d'imputation paramétrique:exemple

Un seul paramètre *R;* 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 l'estimateur par ratio.

Remarque: si les P_k ont été estimés, dans tous les cas la variance 'design based ' du paramètre d'ajustement ℓ est calculable et estimable.

L’estimateur par prédiction: comment ça marche?

Si $\hat{y}_k(x_k;\hat{\beta})$ désigne le prédicteur de y_k dans la loi estimée $\pounds_k$

l’estimateur imputé s’écrit:

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

Sa variance (design based) est facile à calculer car c’est celle de:

$$\sum_r d_k y_k + \sum_o d_k \hat{y}_k(x_k;\beta) + (\sum_o d_k l_k)(\hat{\beta}-\beta)$$

où:

$$l_k = \frac{\partial}{\partial \beta}\hat{y}_k(x_k;\beta)$$

E Exercice: le cas du ratio.

L'estimateur par prédiction:quand est ce que ça marche?

Il fournit un estimateur sans biais ('asymptotique'), pour le total de y .

Il fournit aussi un estimateur sans biais pour le total d'une variable obtenue par transformation 'semi-linéaire' $z_k = c_k y_k$ car:

$c_k \ \hat{y}_k$ est le prédicteur naturel de $c_k y_k$. Typiquement , c_k sera l'indicatrice d'un domaine.

Par contre l'estimation du total de transformations non linéaires de y $(Ex : \mathbf{1}(y \in D$) n'est pas correcte.

L’estimation avec aléa:comment ça marche?

La valeur imputée est tirée dans la loi $\mathcal{L}_k$ estimée. Autrement dit c’est

$\hat{y}_k + e_k$ où e_k est un résidu d’espérance nulle dans la loi estimée.

L’estimateur imputé s’écrit:

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

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

L’estimation avec aléa:comment ça marche?

La valeur imputée est tirée dans la loi $\pounds_k$ estimée. Autrement dit c’est

$\hat{y}_k + e_k$ où e_k est un résidu d’espérance nulle dans la loi estimée.

L’estimateur imputé s’écrit:

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

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

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

où le second terme est celui qui résulte du processus d’imputation.

L’estimation avec aléa: quand est-ce que ça marche?

Au 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*.

Cet accroissement artificiel peut être réduit si on tire les résidus e_k dans une loi jointe ayant les $\mathcal{L}_k$ pour marginales *et présentant des corrélations négatives*, typiquement par des échantillonnages équilibrés (ou, le retour de cube).

Conclusion 1:

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 suivi d'une modélisaton du mécanisme de réponse.

L'imputation par prédiction ne demande que l'estimation d'un prédicteur.
En revanche, l'estimation avec aléa repose sur l'estimation de *la loi* $\mathcal{L}_k$.
Cette procédure est donc moins robuste (plus risquée!) que la première.

Conclusion 2 : quand est-ce que l'imputation ne peut pas marcher DU TOUT?

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 ou de transformations linéaires ou même non linéaires de cette variable.

Il est dangereux pour l'élaboration de statistiques croisant deux ou plusieurs variables car les corrélations sont altérées par l'imputation.

Il est fondamentalement pervers dans toute analyse de nature économétrique, donnant l'illusion d'un enrichissement des données, alors qu'en réalité, il en est un appauvrissement.