

IMPUTATION PAR PREDICTION PARAMETRIQUE ET EQUATIONS ESTIMANTES : UN ESSAI DE MISE EN COHERENCE

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

RÉSUMÉ

Tant qu'on ne considère qu'une unique variable, on a le choix, pour compenser la non-réponse, entre la technique de repondération et des techniques d'imputation. La première est théoriquement presque parfaite, car elle autorise, en particulier, une estimation consistante de la fonction de répartition. Les secondes deviennent indispensables, en pratique, dans le cas d'une non-réponse partielle.

Elles se basent généralement sur des modèles spéculatifs qui prennent parti sur la nature des variables collectées. Dans cet essai, nous allons chercher à ramener le plus possible l'inférence avec données imputées à une inférence basée sur le plan de sondage assisté d'un modèle de réponse. Deux idées majeures sont utilisées. La première consiste à considérer les paramètres utilisés dans l'imputation comme de simples paramètres d'ajustement. La seconde consiste à utiliser systématiquement les équations d'ajustement (ou 'équations estimantes') pour se ramener à une inférence sous le plan. Dans cette optique, on utilise systématiquement le concept de mécanisme de réponse, mais on s'interdit d'utiliser des probabilités de réponse estimées dans les estimateurs ponctuels. Dans certains cas, en utilisant des régressions instrumentales bien choisies, on peut même se dispenser de toute estimation du modèle de réponse.

Les imputations par prédiction déterministe, cependant, sont invalides pour l'estimation de transformations non linéaires de la variable (c'est-à-dire de la fonction de répartition). On a alors recours à des prédictions avec aléas qui ont le défaut d'ajouter une variance parasite à toutes les estimations. La encore, on peut, en se ramenant aux équations estimantes et en utilisant des techniques d'échantillonnage équilibré, diminuer considérablement cette variance parasite tout en assurant une bonne cohérence entre l'imputation et ce que donnerait la repondération.

1. INTRODUCTION : REPONDERATION OU IMPUTATION ?

On s'intéresse à une enquête où les unités sont munies de poids d_k sans biais, ou asymptotiquement sans biais, ou encore éventuellement sans biais sous réserve de validité d'une correction de non réponse globale. Nous supposons donc effectuées les opérations de choix d'estimateur (poststratification, calage,...) préalable à une correction pour non réponse relative à une variable d'intérêt particulière y . En particulier la variance de l'estimateur de totaux $\sum_k d_k y_k$ est une forme quadratique connue $Q(y_U)$ dont on connaît un estimateur $Q^{\wedge}(y_s)$.

Tant qu'on ne considère que cette variable, on sait qu'on a le choix entre une technique de repondération et des techniques d'imputation pour valeur manquante. La première technique est théoriquement presque parfaite, permet d'élaborer toutes les statistiques possibles. De plus, elle ne base l'inférence que sur l'échantillonnage et ce que le statisticien sait du processus de production de son enquête, en particulier sur les facteurs ayant pu influencer la propension à répondre des unités sondées.

Les techniques d'imputation se basent généralement sur des modèles spéculatifs (surtout chez les économistes !) qui prennent parti sur la nature des variables collectées. Dans cet essai, nous allons chercher à ramener le plus possible l'inférence avec données imputées à une inférence basée sur le plan de sondage assisté d'un modèle de réponse. Deux idées majeures sont utilisées. La première consiste à considérer les paramètres utilisés dans l'imputation comme de simples paramètres d'ajustement sans signification 'économique', comme on le fait généralement pour l'ajustement d'un modèle de réponse. La seconde consiste à utiliser systématiquement les équations d'ajustement (ou 'équations estimantes') pour se ramener à une inférence sous le plan. Celles-ci sont fondamentalement de deux natures : des équations de calage, liées au mécanisme de réponse, et des équations de type 'normales', qui assurent l'ajustement des variables d'imputation.

Dans cette optique, on utilise systématiquement le concept de mécanisme de réponse, mais on s'interdit d'utiliser des probabilités de réponse estimées dans les estimateurs ponctuels. On doit donc, d'abord, comprendre ce qui se passerait si on pouvait utiliser la repondération, avant de simuler cela grâce à des imputations, celles-ci n'ayant **jamais** pour but d'ajouter de l'information, mais simplement de boucher des trous.

Cependant l'imputation par prédiction ne permet jamais d'estimer des indicateurs mettant en jeu des transformations non linéaires de la variable d'intérêt comme la fonction de répartition par exemple. Avec un peu de chance, l'imputation avec aléas fournit une façon de surmonter cette difficulté, mais au prix d'une variance artificielle parasite. On peut la réduire en introduisant des contraintes (liées à la création de covariances négatives entre les aléas imputés) qui nécessitent la mise en œuvre d'échantillonnages équilibrés.

2. REPONDERATION ?

Si nous voulons utiliser une technique de repondération, le but est d'obtenir de nouveaux poids $w_k = d_k * b_k$ plus 'représentatifs', c'est à dire où les b_k sont des estimations des inverses de probabilités de réponse. On se limitera essentiellement ici au cas où les pondérations b_k sont de la forme $F_k(\beta) = F(z'_k \beta)$ dépendant d'un vecteur z_k de variables explicatives de la probabilité de réponse et d'un paramètre β de dimension q qui peut être estimé par des équations estimantes de la forme $\sum_r v_k(\hat{\beta}_r) = 0$. Une condition utile pour que ce type de modèle fonctionne bien est qu'une réponse uniforme soit possible comme situation de référence. Formellement, cela signifie la condition suivante :

$$\forall t > 1, \exists \beta_t : \forall k \quad F_k(\beta_t) = t \quad (2).$$

Dans le cas où $F_k(\beta) = F(z'_k \beta)$, cette condition est réalisée dès que la constante figure dans les variables z , ou, ce qui revient au même, s'il existe un vecteur ligne a tel que $az_k = 1$ pour tout k .

Les fonctions v_k de \mathbf{R}^q dans \mathbf{R}^q peuvent être de formes diverses, mais l'utilisation du calage généralisé (Deville(1998, 2000 ou 2002) par exemple), incluse dans le logiciel CALMAR II (Le Guennec, Sautory 2002), semble être la méthode la plus rationnelle et la plus efficace. Elle permet, en particulier, d'utiliser des variables explicatives de la réponse, z_k , connues uniquement sur les répondants, par exemple certaines variables d'intérêt. On traite donc efficacement les cas les plus fréquents de non réponse 'non-ignorable', pour reprendre une terminologie parfaitement détestable. Dans ce cas, on prendra donc $v_k(\beta) = x_k d_k F(z'_k \beta)$, les x_k étant des q -vecteurs de variables sur lesquelles on détient de l'information auxiliaire permettant un calage, et, en principe, expliquant assez bien la variable d'intérêt y .

On tient compte alors de ce que la non-réponse peut être vue comme une phase supplémentaire d'échantillonnage, mais non contrôlée. Pour toute variable y et toute variable dérivée de y , tant linéaire (comme $y_k * \mathbf{I}_k^D$, où \mathbf{I}_k^D est l'indicatrice d'un domaine D) que non-linéaire (comme un fractile, $I(y_k < t)$), 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**, c'est-à-dire, essentiellement, si les variables du modèle de réponse sont bien choisies (même si ce sont des variables d'intérêt). De plus, dans certains cas, 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). La variance de $\hat{Y}_{pond} = \sum_r d_k b_k y_k$ vaut alors $Q(y_U) + E(Q_s(e_s))$ où Q_s est la variance conditionnelle à s et e_s les résidus de la régression de y sur les x utilisant les instruments $F_k'(\beta) = z_k F_k'(z_k' \beta)$. Les principes de l'estimation de variance s'en déduisent (Caron, Deville, Sautory 1998).

On notera que le rôle des variables z est de corriger de biais dus à la non-réponse alors que celui des variables x , qui forment une information auxiliaire dans s sur r , est de réduire la variance conditionnelle de $Q_s(y_s)$ à $Q_s(e_s)$.

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 formellement 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 le cas où les imputations se basent sur des techniques de prédiction paramétriques et on montrera en quoi la façon dont les paramètres (qui sont 'ancillaires' dans ce contexte) sont estimés influe sur les propriétés des estimations d'intérêt : biais et variances d'estimations de totaux.

3. IMPUTATION PAR PREDICTION PARAMETRIQUE

3.1 Imputation sans paramètre : le cold-deck

Les y_k manquant sont les valeurs d'une variable connue g_k . L'estimateur s'écrit, on notant R_k la variable d'espérance $P_k (=b_k^{-1} = F_k(\beta))$ si le modèle de réponse est correct) qui vaut 1 pour les répondants et 0 sinon :

$$\hat{Y}_{imp} = \sum_r d_k y_k + \sum_o d_k g_k = \sum_s d_k (R_k y_k + (1 - R_k) g_k).$$

Son biais vaut $E(\hat{Y}_{imp} - Y) = E(\sum_s d_k ((1 - R_k)(g_k - y_k)))$. Le biais conditionnel à s est donc $\sum_s d_k (1 - P_k)(g_k - y_k)$ et le biais non conditionnel vaut $\sum_U (1 - P_k)(g_k - y_k)$ si on suppose que les P_k ne dépendent pas de s . Pour la variance, on a:

$$Var(\hat{Y}_{imp}) = Var(E(\hat{Y}_{imp}|s)) + E(Var(\hat{Y}_{imp}|s)) = Var(\sum_s d_k (1 - R_k)(g_k - y_k)) + EVar(\sum_s R_k (g_k - y_k)|s).$$

Le dernier terme de cette expression peut être évaluée facilement si on a spécifié le modèle de réponse. Pour un modèle Poissonien, par exemple, c'est $\sum_s P_k (1 - P_k) (y_k - g_k)^2$.

3.2 Imputation par prédiction paramétrique

Comme on peut évaluer sur r les distorsions des g_k par rapport aux y_k , on imagine de les éliminer en introduisant un paramètre à p dimensions réelles θ . Autrement dit on supposera que y_k peut être mieux prédit par une fonction $g_k(\theta)$ pour une bonne valeur du paramètre. Evidemment, en général, les fonctions g_k utiliseront les variables auxiliaires « explicatives » x_k connues sur s et auront une forme déduite d'un modèle de type économique-social plausible. Ceci nous importe peu ici, car nous retenons uniquement la forme analytique des prédicteurs (qui peuvent être communs à une foule de modèles!). Si, par nos relations terrestres ou extra-terrestres, nous pouvons avoir accès à la 'bonne' valeur du paramètre, nous sommes dans le cas du Cold-Deck. Si l'Omniscient refuse de nous donner cette valeur et nous dit de nous débrouiller, nous chercherons à l'estimer à partir des données. En général, et nous nous bornerons ici à ce cas, on utilisera un système de p équations estimantes de la forme:

$$\sum_r u_k(y_k, x_k; \theta) = \sum_r u_k = 0 \quad (3-2-1)$$

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

Exemple 1: On utilise un unique paramètre noté R et $g_k = R x_k$. On n'aura donc qu'une seule équation estimante. Si on utilise par exemple $\sum_r d_k (y_k - R x_k) = 0$ dans le but d'annuler l'erreur portant sur la partie non imputée de l'estimateur, le résultat est la classique imputation par ratio. Si, de plus, le modèle de réponse est un SAS, l'estimateur correspondant n'est autre que l'estimateur par ratio. Mais il y a d'autres choix de l'équation estimante, comme nous allons le voir.

Exemple 2: On prédit y_k par une combinaison linéaire des x_k . Le 'modèle' sous-jacent est donc du type régression et les équations estimantes seront les équations normales de la régression, ce qui laisse le choix entre les moindres carrés ordinaires, les moindres carrés pondérés ou la technique des variables instrumentales !

Exemple 3: La variable y_k est une variable (0-1); son prédicteur est un nombre compris entre 0 et 1 qui, pour certains modèles, s'identifie à la probabilité qu'on ait $y_k = 1$. Par exemple, pour une régression logistique utilisant les x_k comme explicatives on pourra utiliser les équations normales $\sum_r d_k (y_k - f(x_k; \theta))$ où $f(.) = \exp(.) / (1 + \exp(.))$.

3.3 Variance du paramètre d'ajustement

Le choix des équations estimantes détermine les propriétés de l'estimateur obtenu. D'autre part, si les P_k ont été estimées, la variance 'design based' du paramètre d'ajustement θ est calculable et estimable. Partons, en effet, des équations (3-2-1) et notons $\hat{\theta}_r$ sa solution (supposée bien définie, unique et régulière) pour l'échantillon de répondants r . L'espérance de $\sum_r u_k(y_k, x_k; \theta)$ pour le modèle du mécanisme de réponse vaut $\sum_s P_k u_k(y_k, x_k; \theta)$. Notons θ_s la solution de $\sum_s P_k u_k(y_k, x_k; \theta_s) = 0$. On obtient par linéarisation au voisinage de θ_s : $\sum_r u_k(y_k, x_k; \theta_s) + \sum_s P_k \frac{\partial}{\partial \theta} u_k(y_k, x_k; \theta_s) (\hat{\theta}_r - \theta_s) = 0$ de sorte que la variable linéarisée de $\hat{\theta}_r$ (conditionnelle à s , mais c'est un détail pour ce qui nous concerne) est :

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

Dans cette expression $\frac{\partial}{\partial \theta} u_k(y_k, x_k; \theta_s) = u'_k$ désigne la matrice des dérivées partielles de u_k .

On obtient donc la variance conditionnelle à s de $\hat{\theta}_r$ en portant cette expression dans celle de la variance du modèle de réponse (en général un modèle poissonien ou de sondage aléatoire simple éventuellement stratifié) et on l'estime en portant dans l'estimateur de cette variance conditionnelle l'approximation habituelle :

$$-\left(\sum_r \frac{\partial}{\partial \theta} u_k(y_k, x_k; \hat{\theta}_r) \right)^{-1} P_k u_k(y_k, x_k; \hat{\theta}_r).$$

3.4 Biais de l'estimateur imputé

L'estimateur imputé $\hat{Y}_{imp} = \sum_r d_k y_k + \sum_o d_k g_k(\hat{\theta}_r)$ vaut, à une quantité de biais négligeable près, $\sum_r d_k y_k + \sum_o d_k g_k(\theta_s) + (\sum_o d_k g'_k(\theta_s))(\hat{\theta}_r - \theta_s)$. Comme $\hat{\theta}_r$ estime sans biais (asymptotique) θ_s , le biais conditionnel à s est $\sum_s d_k (1 - P_k)(g_k(\theta_s) - y_k)$. Un des principes de la correction pour non-réponse consiste précisément à éliminer le biais conditionnel et donc à inclure dans les équations estimantes de θ (une des coordonnées de (3-2-1)) l'annulation d'un estimateur de ce biais. Il est donc naturel d'utiliser $\sum_r d_k (b_k - 1)(g_k(\hat{\theta}_r) - y_k) = 0$. Il se trouve que cette expression peut se réécrire :

$$\sum_r d_k b_k y_k = \sum_r d_k y_k + \sum_o d_k g_k(\hat{\theta}_r) + (\sum_s d_k g_k(\hat{\theta}_r) - \sum_r d_k b_k g_k(\hat{\theta}_r)) \quad (3-4-1)$$

et, comme le dernier terme est un estimateur sans biais de 0, on peut aussi utiliser l'estimante $\sum_r d_k b_k y_k = \sum_r d_k y_k + \sum_o d_k g_k(\hat{\theta}_r)$, qui exprime que l'estimateur imputé du total de y doit prendre la même valeur que celle de l'estimateur repondéré (virtuel).

Mais il y a plus rusé ! Si on l'égalé à 0 le dernier terme de (3-4-1), on exprime le fait que la somme sur s des prédicteurs est égal à la somme sur r des prédicteurs munis des poids de redressement pour non réponse, ce qui constitue aussi une estimable équation estimante, soit de θ si on connaît β , soit une équation de calage estimante de β si on connaît θ ! Cette dernière façon de voir les choses est sans doute la plus riche pour des raisons qui apparaissent dans la suite.

Ceci posé, une équation estimante n'est susceptible d'éliminer le biais (*pour l'inférence basée sur le plan*) que pour ce qui concerne l'estimation du total de y ; contrairement à ce qui se passe dans l'estimation basée sur un modèle explicatif, on n'estime jamais sans biais le total de y sur un domaine D à moins qu'une équation estimante de θ ne soit de la forme $\sum_r d_k (b_k - 1) \mathbf{1}_k^D (g_k(\hat{\theta}_r) - y_k) = 0$ (ou forme apparentées comme il est dit ci-dessus). Le sous-espace des variables dérivées de y estimables sans biais est donc limité par les équations estimantes qu'on s'autorise à utiliser. En particulier la fonction de répartition de y sera généralement mal estimée.

3.5 Possibilités de choix d'équations estimantes pour le paramètre d'ajustement

Elles sont souvent implicitement définies par les équations estimantes que l'on *choisit*. On dispose de nombreuses possibilités, mais, en tout état de cause, on n'a jamais à privilégier l'estimation 'efficace' de θ , qui n'est qu'un paramètre auxiliaire d'ajustement. Les u_k ne sont donc pas (obligatoirement) dérivées d'une log-vraisemblance où d'un quelconque et arbitraire principe inférentiel. On vient d'en voir une illustration avec la recherche d'un estimateur imputé sans biais pour le plan. On peut aussi avoir d'autres idées plus ou moins fondées quoi que naturelles.

Par exemple, une des coordonnées des u_k pourra être $y_k - g_k(x_k; \theta)$, l'équation estimante exprimant que la somme des prédicteurs sur r doit égaler la somme des valeurs observées. Une variante consiste à vouloir que la partie 'connue' de l'estimateur basée sur les répondants ne varie pas si on remplace les vraies valeurs par les prédicteurs, ce qui conduit à utiliser $d_k (y_k - g_k(x_k; \theta))$.

Ces équations sont des variantes des équations normales de la régression (éventuellement non linéaire) soit : $\sum_r h_k (y_k - g_k(x_k; \theta)) = 0$ où les h_k sont des vecteurs de variables instrumentales qu'on peut choisir à sa convenance. Elles sont définies à une transformation linéaire près, et, par exemple, la constante ou les d_k pourront apparaître comme des combinaisons des h_k .

Si le modèle de réponse n'est pas un sondage aléatoire simple on pourra aussi utiliser des équations de la forme: $\sum_r b_k u_k(y_k, x_k; \theta) - \sum_r \varpi_k(x_k; \theta) = 0$ où $\varpi_k(x_k; \theta) = u_k(g_k(x_k; \theta), x_k; \theta)$ de façon à obtenir la même estimation de θ pour un virtuel estimateur repondéré que pour l'estimateur imputé. En dépit des apparences, ces équations sont bien un cas particulier des équations (3-2-1). Il suffit de poser $u_k^* = b_k u_k(y_k, x_k; \theta) - \frac{m}{n} \sum_s \varpi_k(x_k; \theta)$ pour s'y ramener.

On pourra aussi, par exemple, désirer ne pas modifier l'estimateur du total des y obtenu par repondération et utiliser l'équation estimante $\sum_r d_k b_k y_k - \sum_s d_k g_k(x_k; \theta) = 0$. Sauf cas particulier, ceci ne conduit pas à un estimateur conditionnellement sans biais, ce qui ne réduit que peu l'intérêt de cette façon de faire, surtout si on pense que les prédicteurs ne diffèrent des y_k que par un terme d'espérance nulle sous un modèle ad hoc. En fait, on sera tenté d'utiliser l'estimateur plus 'précis' du total $\sum_s d_k g_k(\hat{\theta}_r)$ ou des totaux de type 'domaine' $\sum_s d_k c_k g_k(\hat{\theta}_r)$ avec $c_k = \mathbf{1}_k^D$.

3.6 Variance de l'estimateur imputé

1-Cas général. Avec les notations déjà introduites, l'estimateur avec valeurs imputées par prédiction, dans le cas paramétrique, va s'écrire $\hat{Y}_{imp, pred} = \sum_r d_k y_k + \sum_o d_k g_k(x_k; \hat{\theta}_r)$. Sa variance (au sens design based) est en principe assez facile à calculer.

En effet, soient $e_k = y_k - g_k(x_k; \theta_s) = y_k - g_k$ les 'résidus vrais' et $\tilde{e}_k = y_k - g_k(x_k; \hat{\theta}_r) = y_k - \hat{g}_r$ les résidus empiriques. L'estimateur s'écrit $\sum_r d_k (g_k + e_k) + \sum_o d_k \hat{g}_k$, soit, approximativement, $\sum_s d_k g_k + \sum_o d_k e_k + (\sum_o d_k l_k)(\hat{\theta}_r - \theta_s)$ avec $l_k = \frac{\partial}{\partial \theta} g_k(x_k; \theta_s)$, noté comme un vecteur ligne. La variance conditionnelle du premier terme est nulle d'où on déduit que la variable linéarisée (pour la variance conditionnelle à s) de l'estimateur vaut: $e_k + (\sum_s (1 - P_k) d_k l_k) \text{lin}(\theta_s)_k$. Pour l'estimation de variance, cette linéarisée sera approximée par $\tilde{e}_k + (\sum_o d_k l_k) \text{lin}(\theta_s)_k$.

Pour obtenir l'estimation de la variance totale il faut ajouter l'estimation de $Var(\sum_s d_k g_k(x_k; \hat{\theta}_s))$. Enfin, pour obtenir l'écart quadratique moyen, il faut encore ajouter le carré du biais conditionnel si on n'a pas choisi l'option qui consiste à l'annuler grâce à une équation estimante.

2-Cas où les équations estimantes sont des équations normales. Le vecteur de paramètres d'ajustement est toujours solution du système d'équations (3-2-1) qui s'écrivent maintenant :

$$\sum_r h_k (y_k - g_k(x_k; \hat{\theta}_r)) = \sum_r h_k \tilde{e}_k = 0 \quad (3-6-2a)$$

La linéarisée de $\hat{\theta}_s$ vaut donc: $\text{lin}(\hat{\theta}_r)_k = t_k = -(\sum_s P_k l_k)^{-1} P_k h_k e_k$. Par suite la variance conditionnelle de l'estimateur est celle de $e_k - (\sum_s (1 - P_k) d_k l_k) (\sum_s P_k h_k l_k)^{-1} P_k h_k e_k$.

Supposons, ce qui ne nous engage à rien pour l'instant, que les instrument soient écrits sous la forme $d_k(b_k - 1)h_k$ au lieu de h_k . Ceci signifie qu'on a choisi d'annuler le biais des composantes de $h_k y_k$. On peut naturellement admettre que le biais de y a été annulé ce qui signifie qu'on dispose d'un vecteur (ligne) v tel que $v h_k = l$ pour tout k . La variance de l'estimateur (conditionnelle à s) est celle de :

$$\sum_r d_k e_k - (\sum_o d_k l_k) (\sum_r d_k (b_k - 1) h_k l_k)^{-1} \sum_r d_k (b_k - 1) h_k e_k,$$

soit, de façon équivalente, (en remplaçant la somme sur o par quelque chose qui a la même espérance) celle de :

$$\sum_r d_k e_k - (\sum_r d_k (b_k - 1) l_k) (\sum_r d_k (b_k - 1) h_k l_k)^{-1} \sum_r d_k (b_k - 1) h_k e_k \quad (3-6-2b).$$

Le tour de magie consiste à vérifier que $(\sum_r d_k (b_k - 1) l_k) (\sum_r d_k (b_k - 1) h_k l_k)^{-1} = v$ et que $v \sum_r d_k (b_k - 1) h_k e_k = \sum_r d_k (b_k - 1) e_k$, de sorte que la variance conditionnelle n'est donc autre que celle de $\sum_r d_k b_k e_k$, c'est à dire que la variable linéarisée (conditionnelle à s) est $d_k e_k$.

Remarque : Comme dans tous les tours de magie il y a une explication simple. Si on remonte à 3.4, on voit qu'on se trouve dans le cas où l'estimateur imputé est identique à l'estimateur repondéré dont la variance conditionnelle est bien connue et donnée par l'astuce des résidus.

3-Estimation de la variance. La variance conditionnelle s'estime donc de façon claire et simple (sous le modèle de réponse). Pour ce qui concerne la partie non conditionnelle, on remarquera que l'espérance conditionnelle de l'estimateur vaut (avec l'approximation linéaire usuelle) $\sum_s P_k d_k y_k + (1 - P_k) d_k g_k(\hat{\theta}_s) = \sum_s (1 - P_k) d_k (g_k(\hat{\theta}_s) - y_k) + \sum_s d_k y_k$. Le premier terme est nul si on est sans biais pour le total de y , par définition de θ_s . On est donc ramené à estimer la variance du second, c'est-à-dire celle de l'estimateur qu'on obtiendrait s'il n'y avait pas de non réponse. Ce problème qui semble théoriquement simple est pratiquement plutôt difficile. Le logiciel POULPE (Caron, Deville, Sautory, (1998)) y apporte une solution satisfaisante et le lecteur est renvoyé à la documentation idoine.

3.7 Cohérence entre l'estimation du modèle de réponse et du prédicteur

1-Choix des auxiliaires. La mise en place de l'imputation est donc conditionnée à la résolution des deux systèmes d'équations, de dimensions respectives q et p :

Des équations de calage pour estimer le modèle de réponse, avec, en général, $F_k(\beta) = F(z'_k \beta)$, z variables explicatives de la probabilité de réponse

$$\sum_r x_k d_k F_k(\beta) = \sum_s x_k d_k \quad (3-7-1)$$

et

$$\sum_r d_k (F_k(\beta) - 1) h_k(y_k - g_k(x_k; \hat{\theta}_r)) = 0 \quad (3-7-2),$$

équations normales permettant d'estimer des paramètres d'ajustement destinées à éliminer des biais générés par l'imputation. Les x sont des variables auxiliaires connues sur s , comme les (en général la !) fonction(s) g_k , les h sont des 'instruments' dont on a besoin que sur r . Sous cette forme générale, il n'y a pas de problème, on calcule β avec (3-7-1) puis θ avec (3-7-2). On rate cependant quelques avantages de cohérence. Ceux-ci sont assez clairs si $p=q$.

2-Cas $p=q$. Si $g_k(x_k; \theta) = x_k' \theta$ (régression linéaire), on a $x_k' = l_k$. Dans le cas général, on sera tenté de garder la même définition, à ceci près que l_k dépend de $\hat{\theta}_r$ et qu'on sera amené à résoudre simultanément les deux systèmes (sauf dans le cas de la régression). Un avantage est que le tour de passe-passe du (3-6-2b) s'explique par une relation exacte au lieu d'une relation valide approximativement.

Mais il y a beaucoup mieux dans la recherche de cohérence en jouant sur le choix des h . Nous supposons explicitement que la condition (2) est réalisée. On a déjà vu qu'une des coordonnées des h devait (presque nécessairement) être la constante l (la variable gratuite). Si nous voulons, maintenant, que l'estimation de θ soit insensible à la valeur de β nous sommes conduit aux équations $\sum_r d_k F_k'(\beta) h_k(y_k - g_k) = 0$. On est conduit à utiliser les $h_k = F_k' / (1 - F_k)$ comme instruments, mais il y en a apparemment un de trop. Erreur, car un nouveau tour de magie apparaît en trois temps:

- *a* - si on centre le problème en β_t avec $t = \sum_s d_k / \sum_r d_k$ (relation (2) + le fait que l fait toujours partie de l'information auxiliaire si on suppose la taille de s connue), on a, en dérivant la relation (2) par rapport à t $F_k' d\beta_t / dt = 1$ ce qui prouve que la constante est combinaison linéaire des coordonnées de F' pour tout k !

Comme en général la valeur de t est immédiate à calculer et que $F_k = t$ les instruments seront simplement les $F_k'(\beta_t)$ et les équations estimantes (3-7-2) deviennent $\sum_r d_k F_k'(\beta_t)(y_k - g_k(x_k; \hat{\theta}_r)) = 0$. On obtient donc des estimateurs imputés débiaisés sans avoir besoin d'estimer le modèle de réponse.

- *b* - Encore plus fort : d'accord on a estimé β , mais avec un modèle où la relation (2) est généralisée de la façon suivante : $\forall \beta \forall t > 0, \exists \beta_t : \forall k \quad F_k(\beta_t) - 1 = t (F_k(\beta) - 1)$ (ça veut dire que les $\logit(P_k)$ changent d'une valeur indépendante de k quand on se promène sur la courbe $t \rightarrow \beta_t$). On a alors $F_k'(\beta_t) d\beta_t / dt = (F_k(\beta) - 1)$ pour tout k et le même miracle se produit sur les équations estimantes du prédicteur : on est sans biais si on les prend de la forme $\sum_r d_k F_k'(\hat{\beta})(y_k - g_k(x_k; \hat{\theta}_r)) = 0$. Si on y regarde de plus près on voit qu'on est même protégé contre les erreurs d'estimation sur β .

- *c* - Dans le cas où $F_k(\beta) = F(z'_k \beta)$, et donc que $F_k' = z_k F'(z'_k \beta)$, et que, de plus, il existe un vecteur ligne a tel que $az_k = l$ pour tout k , on voit immédiatement que :

- dans la situation du *a* ('centrage' du modèle de réponse sur des probabilités égales) les équations estimantes seront tout bonnement $\sum_r d_k z_k (y_k - g_k(x_k; \hat{\theta}_r)) = 0$. Dans la situation du *b*, la condition généralisant (2) implique que le modèle de réponse est un logit, et les équations précédentes garantissent l'absence de biais pour n'importe quelle valeur de β . On a donc plus besoin **du tout** d'estimer le modèle de réponse, mais seulement de sélectionner les bonnes variables (**éventuellement d'intérêt**) sans oublier la constante. Si, éventuellement (pour l'estimation de la variance, par exemple), on a besoin d'estimer les

probabilités de réponse, on pourra le faire en une seule étape si on doit utiliser la valeur estimée du paramètre de prédiction. Cette propriété du logit peut se généraliser si les prédicteurs g_k possèdent la propriété de ratio suivante :

Il existe une fonction réelle strictement croissante et dérivable φ vérifiant :

$$\forall \theta, \exists \delta \in \mathbf{R}^p : \forall k \forall t > 0 \quad g_k(\theta + t\delta) = g_k(\theta) \varphi(t) \quad (3-8).$$

Il est facile de vérifier que cette condition est vérifiée pour beaucoup de prédicteurs utilisés dans la pratique. Par exemple, dans le cas de la régression, on a $g_k(\theta) = x_k' \theta$ et on peut prendre $\delta = x$ et $\varphi(t) = 1+t$.

La propriété désirée (estimateur sans biais avec les équations estimantes de β ci-dessus) résulte du fait qu'on a en dérivant la relation par rapport à t $g_k' \delta = g_k(\theta) \varphi'(t)$ et que g_k est une combinaison linéaire des coordonnées de g_k' .

3-Cas où $p \neq q$.

- a - Si $p < q$, c'est-à-dire qu'il y a plus d'instruments venant du modèle de réponse que de paramètre dans le prédicteur, on cherchera, par exemple, à utiliser un ensemble de combinaisons linéaires permettant de diminuer la variance conditionnelle à s de l'estimateur imputé.

- b - Si $p > q$, c'est-à-dire s'il y a moins d'instruments venant du modèle de réponse que de paramètre dans le prédicteur, on sera libre de choisir des instruments supplémentaires à son goût (les l_k sont toujours disponibles!), ou, éventuellement, on pourra rajouter des variables dans le modèle de réponse, ce qui revient au même quand on est dans la situation décrite en 3-8-2-c ci-dessus.

3.9 Résumé et exemples

Il y a interdépendance entre les équations estimantes du modèle de réponse, qui assurent les moyens d'éviter des biais dus à la non-réponse et celles que l'on choisit pour les imputations. En particulier, pour éviter des biais dus à l'imputation il s'avère opportun d'utiliser des variables instrumentales identiques ou proches de celles du modèle de réponse. Inversement les variables de calages qui interviennent dans l'estimation du modèle de réponse doivent être proches de celles qui interviennent comme 'explicatives' dans le modèle de prédiction car elles assurent une minimisation de la variance conditionnelle pour l'estimation du total de la variable d'intérêt. Ceci conduit à recommander d'utiliser des modèles de réponse vérifiant la relation (2) (logit par exemple) et des prédicteurs vérifiant la relation (3-8) (linéaires-régression ou logit) et les équations estimantes cohérentes (et assez symétriques) :

$$\text{Equations de calage :} \quad \sum_r d_k g_k'(\theta) F_k(\beta) = \sum_s d_k g_k'(\theta) \quad (3-9-1)$$

et

$$\text{Equations estimantes 'normales' :} \quad \sum_r d_k F_k'(\beta) y_k = \sum_r d_k F_k'(\beta) g_k(\theta) \quad (3-9-2)$$

Nous allons voir comment cela fonctionne dans trois cas particuliers d'intérêt assez général.

Exemple 1: Imputation par ratio (ou par ratio par classe). Tout est assez bien connu, mais il est utile de comprendre comment cela fonctionne dans la logique que nous venons de mettre en place.

En version de base, on a une réponse uniforme avec calage sur 1, et un modèle de prédiction par règle de trois : l'instrument $z_k = 1$, le prédicteur scalaire x_k nous conduisent aux classiques équations :

$$\sum_r d_k \beta = \sum_s d_k \quad \text{et} \quad \sum_r (y_k - x_k' \theta) = 0 \quad \text{et tout est simple.}$$

En version avec option, ça peut devenir plus coquin (et faire de jolis exercices pour les étudiants !). On peut disposer d'une variable scalaire z_k qui influe sur la réponse (la taille d'une entreprise par exemple). Comme il ne faut pas oublier la constante, on posera donc $b_k = F(\alpha + z_k' \beta)$. Mais du coup on a une surparamétrisation pour l'équation de prédiction. Si le modèle est logit, par exemple, on serait tenté de faire, pour ne pas estimer le modèle de réponse $\sum_r z_k (y_k - x_k' \theta) = 0$ et $\sum_r (y_k - x_k' \theta) = 0$. Sinon, estimer le modèle de réponse et utiliser 'la' bonne équation de prédiction.

Exemple 2: Cellules de réponse et prédiction linéaire par régression.

Ici z_k est l'indicatrice de cellule (q cellules donc), et le modèle de réponse est estimé par : $\sum_r x_k d_k z_k' \beta = \sum_s x_k d_k$ soit en sous-indiquant par i ce qui est relatif à la cellule i les q équations $\sum_{ri} x_k d_k \beta_i = \sum_{si} x_k d_k$. Si x est aussi une variable catégorielle indicée par j , en notant X la matrice des $\sum_{rij} d_k$ et X_s le vecteur des $\sum_{si} d_k$, β est solution de $X\beta = X_s$. Si les x sont identiques aux z (e qui n'a aucune raison d'être en général !), X est diagonale et on a la banale 'postratification formelle'.

Pour la prédiction, les équations normales de la régression sont celles de la régression instrumentale : $\sum_r d_k z_k (y_k - x_k' \hat{\theta}_r) = 0$ soit les équations $\sum_{ri} d_k (y_k - x_k' \hat{\theta}_r) = 0$.

On a implicitement supposé ici, pour la simplicité, que $p=q$. Si ce n'est pas le cas, on voit bien la marche à suivre : si $q < p$ sélectionner de 'bonnes' combinaisons linéaires des x comme instruments dans l'équation de réponse, si $p < q$, sélectionner de 'bonnes' combinaisons linéaires des z comme instruments dans l'équation de prédiction.

Exemple 3 : Le modèle de réponse est logit : $P_k = \exp(z_k' \beta) / (1 + \exp(z_k' \beta))$ et donc $F_k(\beta) = 1 + \exp(z_k' \beta)$ tout bonnement. La variable y est un ratio (financier par exemple) ou une probabilité modélisée également par une régression logistique $y_k = g_k(\theta) = \exp(x_k' \theta) / (1 + \exp(x_k' \theta))$.

Comme le modèle de réponse est logit, il y a miracle et θ s'estime en résolvant les équations normales $\sum_r d_k z_k (y_k - g_k(\hat{\theta}_r)) = 0$. S'il n'y a pas assez d'instruments, comme d'habitude on ajoute ce qu'il faut à partir des x . Ensuite, si besoin est, on résout les équations de réponse en utilisant les instruments $g_k'(\hat{\theta}_r) = x_k' g_k(\hat{\theta}_r) (1 - g_k(\hat{\theta}_r))$ qui dépendent de l'estimation de θ . Or ce besoin existe dès que l'on veut calculer ou estimer la variance puisque les $\sum_s d_k g_k'(\theta)$ jouent le rôle d'une information auxiliaire sur laquelle on se cale.

4. IMPUTATION PARAMETRIQUE AVEC ALEA

4.1 Généralités et position du problème

Les y_k manquant sont considérés comme des variables aléatoires dont on a estimé la loi \mathcal{L}_k à partir de l'échantillon r de répondants. Nous nous limiterons ici au cas où l'espérance de la loi a été estimée de façon paramétrique conformément à ce qui a été dit dans la partie 3. L'imputation consiste à ajouter des aléas aux espérances. La loi de ces aléas dépend de la nature des données, et, éventuellement de paramètres supplémentaires estimés sur les répondants. 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).

Le but de l'opération est d'obtenir une estimation consistante de la fonction de répartition de y et donc de toute estimation de total d'une transformation non-linéaire de y . *Si la réalité était conforme au modèle*, on peut montrer que ce but serait à peu près atteint ; par contre, on ne peut donner aucune justification basée sur l'échantillonnage (y compris le modèle de réponse) de ce fait.

On distinguera ici entre des variables binaires (ou 0-1) –section 4.2-, dont la loi ne peut être qu'une Bernoulli, les variables discrètes (ou qualitatives) –section 4.3- et les variables continues (ou numériques ou réelles) –section 4.4. Le cas de variables discrètes ordonnées (entières 'petites' comme des tailles de ménage par exemple) se ramène, selon les circonstances et le modèle, à l'un des deux derniers cas. Dans tous les cas on peut considérer que l'imputation avec aléa consiste à ajouter une variable aléatoire centrée aux prédicteurs (pour une variable qualitative on considérera le vecteur des indicatrices des modalités, comme d'habitude) soit formellement $\tilde{y}_k = \hat{g}_k + e_k^*$.

La loi des e_k^* 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{g}_k (qui ne dépend que de r), mais une variance supplémentaire 'parasite' affecte l'estimateur, y compris pour les quantités qui sont estimées sans biais par des prédicteurs. Celle-ci est en général loin d'être négligeable (un exercice classique montre qu'elle ajoute souvent 10 à 20% à la variance d'échantillonnage), surtout si les imputations se font de façon indépendante. En notant E^* et V^* l'espérance et la variance de la loi des imputées, on a naturellement (comme dirait Chirac) : $Var(\sum_r d_k y_k + \sum_o d_k \tilde{y}_k) = Var(\sum_r d_k y_k + \sum_o d_k \hat{g}_k) + Var^*(\sum_o d_k e_k^*)$, c'est à dire la variance de l'estimateur par prédiction plus la variance 'parasite' due à l'imputation. Si les imputations sont indépendantes, le dernier terme (vu comme séquentiel) se comporte comme une marche aléatoire et sa variance vaut $\sum_o d_k^2 \sigma_k^2$ avec une notation évidente pour variance de la loi des aléas. Le but que nous nous proposons est de réduire cette variance qui est de l'ordre de grandeur de $Card(o)$ à une quantité bornée que soit la taille de o

.On y arrive , techniquement, en introduisant des covariances négatives entre les imputées. En voyant les choses séquentiellement, la suite des sommes partielles de $\sum_o d_k e_k^*$ va se comporter comme un processus borné en probabilité par une loi unique, de la nature d'un pont Brownien (cas des 4.2 et 4.3) ou d'un processus auto régressif (4.4).

4.2 Imputation d'une variable binaire par échantillonnage équilibré

On suppose que les 'prédicteurs' Q_k ont été estimés de façon consistante par des équations estimantes de type normales : $\sum_r h_k (y_k - Q_k(x_k ; \hat{\theta}_r)) = \sum_r h_k \tilde{e}_k = 0$. L'imputation va consister à attribuer des valeurs 0 ou 1 aux k de o . Imputations indépendantes signifie réaliser un échantillonnage Poissonien des unités imputées à 1, recette connue pour sa faible efficacité. On aurait plus confiance en un échantillonnage de taille fixe (quitte à s'arranger pour que la somme des Q_k soit arrondie à un entier), ou, plus généralement, à introduire des contraintes dans l'échantillonnage. Si les h_k sont disponibles aussi sur o , il sera naturel d'introduire les contraintes $\sum_o h_k \tilde{y}_k = \sum_r h_k Q_k$ ce qui revient à réaliser un échantillonnage équilibré dans les canons de la méthode du Cube (Deville, Tillé(2004), enfin !). Un avantage formel sympathique est que l'estimation des Q_k sur les données imputées sera la même qu'à partir des répondants.

Un autre avantage, que nous ne ferons qu'indiquer sans l'exploiter de façon plus approfondie, réside dans une réduction intéressante de la variance d'imputation. Remarquons d'abord que, si (une des coordonnées des h_k vérifie) $h_k = d_k \omega_k$ ou ω est une variable d'intérêt (éventuellement elle-même imputée !), la variance d'imputation de l'estimateur du total des $\omega_k y_k$ (des ω sur le domaine $y=1$ donc) est nulle. De façon générale, (Deville, Tillé(2005)) la variance d'imputation de l'estimateur du total $\sum_s d_k y_k \omega_k$ sera celle des résidus des $d_k \omega_k$ sur les h_k (quelle que soit la façon dont on les a choisis, ce qui ouvre des perspectives...).

4.3 Imputation équilibrée d'une variable qualitative

C'est sensiblement la même chose. On suppose les Q_{ki} , probabilités pour que M^ossieu k soit dans la position i de la variable qualitative estimées de façon consistante (par exemple à partir d'un ajustement log-linéaire $Q_{ki} = \exp(x_k' \theta_i - c_k)$ réalisé sur les répondants avec les mêmes exigences qu'au § 3). La encore, on est amené à échantillonner des 1 dans un tableau $o \times I$ de façon à ce que la somme en ligne soit égale à 1. Quant à la somme en colonne, on peut essayer de la contrôler de façon à ce que ses totaux n_i soit égaux aux $\sum_{k \in o} Q_{ki}$ (si ces quantités ne sont pas entières on peut admettre qu'on les a arrondis par une procédure de raking-ratio initiale). Ceci peut se faire (si $Card(I)$ n'est pas trop gros) en utilisant un échantillonnage de Poisson multidimensionnel tel que décrit dans Deville(2005), ou, à défaut, l'heuristique de raking-ratio suivante :

-Étape k : imputer i à k avec les probas Q_{ki}

-Mettre à jour les totaux : $n_i \rightarrow n_i - I$, les autres totaux ne changent pas

-Ajuster les probas Q_{ki} par raking-ratio sur les nouveaux totaux pour les lignes à partir de $k+1$.

-Étape $k+1 \dots$

Ceci dit, on peut faire plus fort. Comme au 4.3 on peut avoir envie de contraindre un peu plus les imputations par des choses du genre $\sum_{k \in o} h_k \tilde{y}_k = \sum_{k \in o} h_k Q_k$ en notant comme un vecteur ligne les vectorialisées des \tilde{y}_k et des Q_{ki} (si vous avez bien suivi, l'égalité précédente est une égalité entre matrices qui contraint l'échantillonnage). La encore la méthode du Cube peut s'appliquer pour réaliser une imputation équilibrée, avec les mêmes avantages qu'au 4.3 : stabilité des estimations du modèle d'imputation, annulation ou réduction de la variance d'imputation des estimateurs de certains totaux.

4.4 Imputation d'une variable quantitative à l'aide d'une surmartingale positive

a -Les variables quantitatives permettent de déployer plus facilement l'*arsenal probabiliste*. On impute donc des $\tilde{y}_k = \hat{g}_k + aléas$. Les aléas seront pris de la forme $\sigma_k e_k^*$ où les e_k^* suivent la même loi de probabilité, postulée par le modèle d'imputation, et où les σ_k sont des variables connues sur tout s . Ceci peut prendre plusieurs formes, allant de la gaussienne de service à une loi estimée sur les répondants. A titre indicatif, j'aime bien la procédure suivante. On s'intéresse aux erreurs de prédiction empiriques \tilde{e}_k pour k dans r telles qu'on les

récupère dans une estimation du genre 3-6-2a . Si on en a les moyens on essaye d'estimer une fonction de variance de la forme $\sigma_k^2 = x_k' \tau$ par une méthode de type 'moments' en résolvant $\sum_r h_k \tilde{e}_k^2 = \sum_r h_k x_k' \tau$ avec des instruments h de même dimension que τ (par exemple le cas où x est de dimension 1 est classique, comme celui où c'est la variable constante et gratuite 1). Puis on fabrique un stock de résidus $e_k^* = \tilde{e}_k / \sigma_k$ qui sont un échantillon de la loi postulée génératrice des écarts. Ensuite, s'ils ont une bonne tête, on ajuste une loi du commerce, sinon (ce que je préfère), on les considère comme une estimation non-paramétrique de la dite loi.

Pour l'imputation, la méthode 'naïve' consiste à tirer indépendamment des e_k^* pour k dans o et à imputer $\tilde{y}_k = \hat{g}_k + \sigma_k e_k^*$ et on a $Var^*(\sum_r d_k y_k + \sum_o d_k \tilde{y}_k) = Var^*(\sum_o d_k e_k^*) = \sum_o d_k^2 \sigma_k^2$. Si on fait le travail séquentiellement, la variance 'courante' de $\sum_{k=1}^i d_k e_k^*$ vaut $\sum_{k=1}^i d_k^2 \sigma_k^2$ et augmente linéairement. Il en va de même de sommes du genre $S_i = \sum_{k=1}^i h_k e_k^*$ pour un vecteur d'instruments arbitraires h_k dont la (matrice de) variance augmente comme $\sum_{k=1}^i h_k h_k' \sigma_k^2$.

On va montrer qu'on peut 'astucieusement' contrôler les S_i à rester des $O_p(1)$ au lieu d'être des $O_p(i^{1/2})$ et à garder une variance bornée quel que soit l'effectif de o . En particulier une des coordonnées des h sera, naturellement (comme dit ...), $d_k \sigma_k$ de façon à retrouver, à peine altérée, l'estimation par prédiction du total des y . Plus généralement, si les prédicteurs ont été ajusté par des équations du genre 3-6-2a, le fait de reprendre les mêmes instruments assure que la solution des équations estimantes de θ sera la même à partir des données imputées qu'à partir des données sur les seuls répondants (à condition, naturellement, que les instruments soient aussi disponibles sur s). Avant de montrer comment le faire, on va donner une autre bonne raison de le faire basée sur l'idée de réduire et le biais et la variance dus à l'imputation.

b- Supposons qu'on veuille estimer $\Phi = \sum_U \varphi(y_k)$ par $\hat{\Phi} = \sum_r d_k \varphi(y_k) + \sum_o d_k \varphi(\tilde{y}_k)$. Pour une fonction φ régulière et des aléas pas trop gros on aura approximativement :

$$\hat{\Phi} = \sum_s d_k \varphi(\hat{g}_k) + \sum_r d_k \varphi'(\hat{g}_k) \tilde{e}_k + \sum_o d_k \varphi'(\hat{g}_k) e_k^* + \sum_r d_k \varphi''(\hat{g}_k) \tilde{e}_k^2 + \sum_r d_k \varphi''(\hat{g}_k) e_k^{*2}.$$

En prenant l'espérance on constate que le biais d'imputation (si le modèle est correct) est sensiblement éliminé puisque $E^* \hat{\Phi} \approx \sum_s d_k \varphi(\hat{g}_k) + \sum_s d_k \varphi''(\hat{g}_k) \sigma_k^2$. La variance d'imputation, soit $Var^*(\sum_o d_k \varphi'(\hat{g}_k) e_k^*)$, va aussi se trouver réduite; d'abord, s'il se trouve que (par chance!) on ait une régression exacte $\varphi_k = \varphi'(\hat{g}_k) = \gamma h_k$ pour un vecteur ligne γ , la variance est nulle. Sinon, soit Σ_e la matrice des variances-covariances des e^* . C'est une matrice symétrique positive de dimension $Card(o)$ dont le noyau contient les vecteurs colonnes des composantes des h . La variance d'imputation vaudra (avec φ vecteur des φ_k) $\varphi' \Sigma_e \varphi$. Si les e_k^* sont gaussiens cette quantité peut être évaluée. En effet, en cas d'indépendance la matrice de variance des e_k^* n'est autre que $\Sigma = diag(\sigma_k^2)$. Dans le cas contraint (il y a des méthodes pour le faire exactement et on admettra que celle qui va être proposée au c ci-dessous le fait approximativement), la matrice est gaussienne conditionnée par $\sum_o h_k e_k^* = 0$ et vaut donc, avec H matrice des h_k' empilés, $\Sigma (Id - H(H' \Sigma^{-1} H) H' \Sigma^{-1})$ de sorte que la variance de $\sum_o d_k \varphi_k e_k^*$ n'est autre que celle des résidus des $d_k \varphi_k$ régressés sur les h_k avec les poids σ_k^{-2} . Bref encore le truc des résidus.

Remarque : Ces résultats ne permettent pas directement de parler de la fonction de répartition car les fonctions $\varphi_i(y) = I(y < t)$ ne sont pas assez lisses. Avec un peu de travail, on peut cependant montrer que presque tout ce qui vient d'être dit reste valide dans le cas de la fonction de répartition.

c- Et donc maintenant comment le faire ?

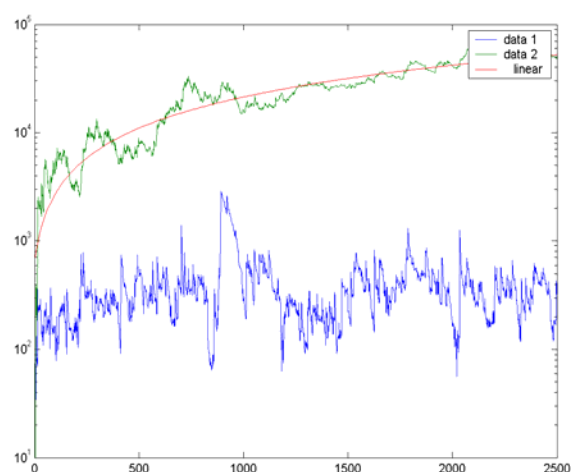
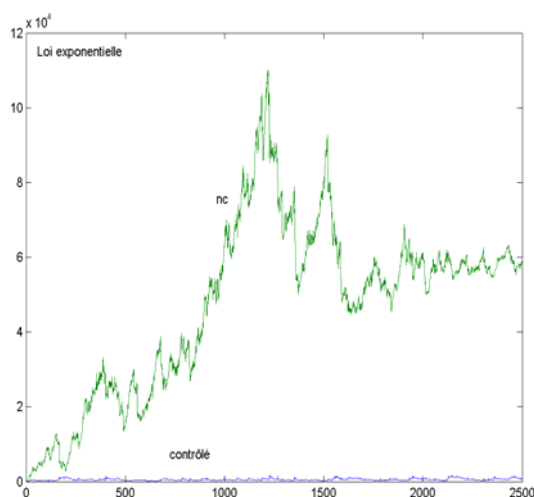
L'algorithme est très simple. On examine le produit scalaire $S_i' h_{i+1}$. S'il est négatif on tire e_{i+1}^* dans la partie négative de la loi des e^* , dans la partie positive dans le cas contraire. De ce fait la norme de S_i décroît en espérance dès qu'elle a dépassé un certain seuil.

Voici une démonstration dans le cas simplifié où les e^* ont une loi symétrique et sont de variances égales, et où les h_k ont tous la même norme : on a $S_i = S_{i-1} + h_i e_i^*$ et donc $\|S_i\|^2 = \|S_{i-1}\|^2 + \|h_i\|^2 e_i^{*2} - 2|h_i' S_{i-1} e_i^*|$. En prenant l'espérance conditionnelle à S_{i-1} on a $E(\|S_i\|^2 | S_{i-1}) = \|S_{i-1}\|^2 + \|h\|^2 \sigma^2 - 2\|S_i\| E(|u' h_i| | e_i^*)$ où u est le vecteur unitaire de la direction de S . En notant a la borne inférieure en u de l'espérance du dernier terme (qui est toujours strictement positive) on trouve que $E(\|S_i\|^2 | S_{i-1}) \leq \|S_{i-1}\|^2 + \|h\|^2 \sigma^2 - 2\|S_i\| a$ qui est négatif dès que $\|S_i\| > \|h\|^2 \sigma^2 / 2a$. Tant que S reste dans la boule (dite 'de sécurité'), on ne s'inquiète pas; dès qu'il en sort, la

suite de variables aléatoires $\max(\|S_i\| - \|h\|^2 \sigma^2 / 2a, 0)$ est une surmartingale positive qui donc tend presque sûrement vers zéro. En particulier, la variance des S_i est uniformément majorée par celle de la loi de première sortie de la 'boule de sécurité' $\|S_i\| \leq \|h\|^2 \sigma^2 / 2a$.

S'il se trouve que les $h_i e_i^*$ suivent des lois de Gauss sphériques on trouve facilement que le rayon de la boule de sécurité vaut $\sqrt{\frac{\pi}{2}} p \sigma$ où p est le nombre de dimensions des h , et σ , naturellement, l'écart-type des e^* .

Tous ces résultats sont, naturellement, confirmés par de jolies simulations dont voici quelques exemples. La courbe du dessus décrit l'évolution de la norme du vecteur S quand on fait des imputations indépendantes, celle du dessous ce qui se passe quand on utilise la méthode à élastique de rappel décrite ci-dessus. La courbe de droite est en échelle logarithmique, la suite des aléas est la même pour les courbes d'un même graphique, ce qui se repère bien car les grands sauts sont les mêmes dans les deux courbes (parfois en sens inverse, naturellement!).



5. CONCLUSIONS

5.1- Dans tous les cas, la mise au point de l'estimateur imputé, y compris l'évaluation et l'estimation de sa variance, ne peuvent que difficilement faire l'économie d'une étude soignée et 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 \mathcal{L}_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 : total ou totaux fonction de la variable imputée. L'imputation avec des prédicteurs permet aussi l'estimation sans biais de certaines transformations linéaires de la variable imputée (en pratique certains 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) mais, naturellement, 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 appauvrisait.

5.2- Ce papier aura pu paraître un peu abondant et difficile pour le lecteur pris à froid. C'est vrai qu'il résume une demi-douzaine d'articles potentiels et deux chapitres du bouquin que je désespère d'écrire vraiment un jour. J'espère pourtant que les idées émises ici fructifieront. Il suffirait qu'un vague anglophone s'en empare

et se les attribue. Quand on voit, cependant, qu'ils ont encore dix ans de retard dans la compréhension des techniques de calage, on peut penser qu'il y en a encore pour un moment .

Ce papier est sans doute le dernier que j'écris pour un colloque francophone de sondages, sans regret. La statistique m'aura certes bien amusé et fait faire des rencontres. Elle m'aura surtout permis de gagner honorablement ma croûte. Quand je serai payé à ne rien faire, naturellement, je m'amuserai à autre chose, n'ayez pas peur.

RÉFÉRENCES

- Caron, N., Deville, J.C., Sautory, O. (1998) « Estimation de précision de données d'enquêtes : le logiciel POULPE », *Document de travail série méthodologie N°9806*, INSEE, Paris
- Deville, J.C. (1998), « La correction de la non-réponse par calage ou par échantillonnage équilibré », *Actes de la rencontre annuelle de la Société de Statistique du Canada*, Sherbrooke, Canada.
- Deville, J.C. (2000), « Generalized Calibration and Application to Weighting for Non-response », *Actes du Colloque COMPSTAT 2000*, Utrecht, Pays-Bas :Springer
- Deville, J.C. (2002), « La correction de la non-réponse par calage généralisé », *Journées de Méthodologie Statistique de L'INSEE*, Paris, France sur <http://jms.insee.fr>
- Deville, J.C. , Tillé, Y. (2004), « Efficient balanced sampling: The cube method », *Biometrika*, 91,4, p. 893-912.
- Deville, J.C. , Tillé, Y. (2005), « Variance approximation under balanced sampling », *Journal of Statistical Planning and Inference*, 128, p. 569-591
- Haziza, D, (2002) « Inférence en présence d'imputation : un survol », *Journées de Méthodologie Statistique de L'INSEE*, Paris, France sur <http://jms.insee.fr>
- Haziza, D, (2005) « Inférence pour des domaines en présence de données imputées », *Séminaire de Statistique de l'ENSAI*, 18 mars 2005, Bruz, France sur <http://www.ensai.fr>
- Haziza, D, Rao, J.N.K, (2005) « Approche par modèle de non-réponse pour l'inférence en présence de données imputées », *Journées de Méthodologie Statistique de L'INSEE*, Paris, France sur <http://jms.insee.fr>
- Le Guennec, J. , Sautory, O. (2002) « Une nouvelle version de la macro CALMAR de redressement d'échantillon par calage », *Journées de Méthodologie Statistique de L'INSEE*, Paris, France sur <http://jms.insee.fr>
- Neveu, J. (1973), *Martingales à temps discret*, Paris: Masson.