LA CORRECTION DE LA NON-REPONSE PAR CALAGE GENERALISE

Jean-Claude DEVILLE

Laboratoire de Statistique d'Enquête,

Campus de Ker-Lann, 2 rue Blaise Pascal – 35170-BRUZ

-deville@ensai.fr

Le calage ‘traditionnel’

On veut définir de nouveaux poids de la forme

$$w_k = d_k \; F\,(q_k x_k \lambda)$$

où λ est un paramètre d’ajustement à p dimensions.

X est un vecteur de totaux connus sur lequel on ajuste ce paramètre .

$$X = \sum_s d_k \; x_k \; F\left(q_k x_k^{'} \lambda\right)$$

La variance du nouvel estimateur est donnée par le ‘truc’ des résidus:

$$Var\left(\sum_s w_k y_k\right) = Var_{HT}\left(\sum_s w_k e_k\right)$$

Le calage généralisé:

On veut définir de nouveaux poids de la forme:

$$w_k = d_k \; F_k(\lambda)$$

où λ est un paramètre d'ajustement à *p* dimensions.

X est un vecteur de totaux connus sur lequel on ajuste ce paramètre .

$$X = \sum_s d_k \; x_k \; F_k(\lambda) = \sum_s d_k \; x_k \; (1 + z_k^{'} \lambda + ...) \quad = \sum_s d_k x_k \; F(z_k^{'} \lambda)$$

Ce sont les équations de calage .Les poids prennent la forme:.

$$w_k = d_k \; (1 + z_k^{'} \lambda + O(\|\lambda\|^2))$$

Avec:

$$\lambda = \left(T_{szx}^{'}\right)^{-1}\left(X - \hat{X}\right) + O\left(\left\|X - \hat{X}\right\|^2\right) \qquad \text{et} \qquad T_{szx} = \sum_s d_k z_k x_k^{'}$$

Dans le cas linéaire:

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

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

où $\tilde{\beta}$ est solution de :

$$\sum_s d_{kz} z_k \left(y_k - x_k' \beta\right) = 0$$

β est le vecteur des coefficients de la régression instrumentale (*Fuller (1987)* par exemple)) utilisant les z_k comme instruments.

On constate les faits suivants :

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

-- La variance de l'estimateur se calcule en utilisant la technique 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 (en utilisant les résidus empiriques dans un logiciel du genre POULPE).

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

Exemples:

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 :

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

Et les résidus valent: $y_k - \hat{R} x_k$

2 : Estimateur par régression pondérée

Les instruments sont : $z_k = q_k x_k$.

3 : Estimateur par régression optimal (Montanari (1987))

4 : etc…

Non –réponse:modèle de réponse

Le mécanisme de réponse est modélisé par un plan de sondage $q(r;\beta)$ où β est un paramètre inconnu de R^P .

Ce modèle fournit des poids d'extrapolation ‘à la Horvitz-Thompson’:

$$\pi_k^{-1} = F_k(\beta)$$

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

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

Estimation du modèle de réponse

(cas d'une enquête exhaustive)

On va hardiment adopter un principe de calage:

$$\sum_r F_k(\hat{\beta})\, x_k = \sum_U x_k \quad \text{ou} \quad \text{avec} \quad \text{un} \quad GLM :$$

$$\sum_r F\,(z_k^{'}\hat{\beta})\, x_k = \sum_U x_k$$

On peut en donner l'interprétation suivante :

$$X = \sum_r x_k F_k(\beta)\, G_k(\lambda)$$

avec:

$$G_k(\lambda) = \frac{F_k(\beta+\lambda)}{F_k(\beta)} \qquad (et\ on \quad a \quad bien\ G_k(0) = 1)$$

Les G_k sont des fonctions de 'calage généralisé' !!

Estimation du modèle de réponse

(suite)

- $\hat{\beta} = \beta + \lambda$ est un estimateur de β.

- Si, par hasard, on dispose d'une autre estimateur $\hat{\beta}_0$ de β, on peut écrire $\hat{\beta} = \beta + \left(\hat{\beta}_0 - \beta\right) + \lambda$ et l'interprétation est la même.

- On n'a besoin de connaître les F_k (ou les z_k en pratique) que pour les répondants.

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

Des exemples :

Redressement par ratio:

x_k est une variable positive, et $z_k=1$. Autrement dit les réponses manquent au hasard et on cale sur le total X des x_k .

Alors $\hat{Y} = X \frac{Y_r}{X_r}$ où Y_r et X_r sont les totaux sur les répondants .

(il s'agit de la théorie de la règle de trois)

Avec le modèle de réponse de Poisson on obtient
la variance estimée suivante :

$$\frac{X}{X_r}\left(\frac{X}{X_r}-1\right)\sum_r (y_k - R\,x_k)^2$$

Si $x_k = 1$ on obtient :

$$\frac{N}{n}\left(\frac{N}{n}-1\right)\sum_r (y_k-\bar{y})^2$$

$$=\frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{\sum_r (y_k-\bar{y})^2}{n}$$

ce qui est assez naturel.

Un exemple pratique plus original :

- x_k est le vecteur des indicatrices d'une variable qualitative à I modalités $i = 1$ à I. L'effectif N_i de la modalité i est connu.
- z_k est le vecteur des indicatrices d'une autre variable qualitative de même dimension indicé par a .

On pourra voir i comme la CSP au RP (quel jargon!) et a comme la CSP à l'enquête.

On posera $Pr(k \text{ répond}) = P_a = 1/(1+\beta_a)$ si k est classé dans la catégorie a. *(modèle des cellules homogènes de réponse)*

(**suite)**

Les paramètres du modèle de réponse sont estimés par les équations ‘de type calage’ suivantes :

$$N_i = \sum_a R_{ia}\left(1+\hat{\beta}_a\right)$$

Sous l’hypothèse d’un mécanisme de réponse Poissonnien, la variance de l'estimateur calé vaut:

$$\sum_a P_a\left(1-P_a\right)\sum_i \sum_{k\in U_{ai}}\left(y_k-\bar{Y}_i\right)^2$$

Elle est estimée par:

$$\sum_a \hat{\beta}_a\left(1+\hat{\beta}_a\right) \sum_i \sum_{k\in r_{ia}}\left(y_k-\hat{\bar{Y}}_i\right)^2$$

avec

$$\hat{\bar{Y}}_i = \frac{1}{N_i}\sum_a (1+\hat{\beta}_a)\sum_{r_{ia}} y_k$$

Non-réponse après échantillonnage

Dans la logique de ce qui a été dit, on est conduit à estimer β par les équations estimantes de calage suivantes :

$$X = \sum_r x_k d_k F_k(\hat{\beta}) = \sum_r x_k d_k F(z_k' \hat{\beta})$$

On peut réécrire ces équations sous la forme d'authentiques équations de calage (généralisé) :

$$X = \sum_r x_k d_k F_k(\beta)\, G_k(\lambda)$$ avec $$G_k(\lambda) = \frac{F_k(\beta + \lambda)}{F_k(\beta)}$$

où β est la vraie valeur du paramètre .

(Suite)

Les conclusions sont donc les mêmes dans ce contexte :

- $\hat{\beta} = \beta + \lambda$ est un estimateur de β

- Si, par hasard, on dispose d'une autre estimateur de β l'interprétation est la même car λ et l'erreur d'estimation sont de l'ordre de *1/n*.

- On n'a besoin de connaître les F_k (ou les z_k en pratique) que pour les répondants .

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

(suite 2)

Ces résidus seront donc calculés de la façon suivante : on écrit les équations normales de la régression dans U , soit:

$$\sum_U z_k \, (y_k - x_k' B) = 0.$$

L'estimation à partir de r est donnée par la solution des équations normales estimées :

$$\sum_r d_k F_k(\beta) \, z_k \, (\, y_k - x_k' \hat{B}_0 \,) = 0$$

$\hat{B}_0$ est l'estimateur de B qu'on aurait si les vraies probabilités de réponse étaient connues. On va donc calculer:

$$\sum_r d_k F_k(\hat{\beta}) \, z_k (\underbrace{y_k - x_k' \hat{B}}_{résidus}) = 0$$

Conclusions

-Une extension de la méthode bien connue de calage mettant en jeu deux ensemble de variables de même dimension *x* et *z*.

-Estimateurs dont la variance est celle de l'estimateur de Horvitz-Thompson appliquée aux résidus de la régression de la variable d'intérêt sur les *x* en utilisant les *z* comme variables instrumentales. Les *z* apparaissent ainsi comme une généralisation des poids de régression utilisés dans la méthode d'origine.

- Elle est donc efficace surtout quand les deux ensembles de variables sont bien corrélés, et, de ce fait, peut sembler être d'une utilité douteuse puisqu'on peut toujours prendre *z=x* .

Utilité:

-*Les x* sont des mesures entachées d'erreurs (sans biais).

-Estimation par la méthode de partage des poids avec information auxiliaire (voir *Lavallée(2002))*.

-Corriger les biais dus à la non-réponse en utilisant des méthodes de calage, ce qui était l'objet de cette étude.

Avantages:

-La taille du problème est réduite à celle des vecteurs de calage x. Ces variables ont pour utilité, comme dans la méthode classique de réduire la variance de l'estimation.

-Les variables z ont pour fonction de réduire les biais.

-ELLE PEUVENT ÊTRE DES VARIABLES D'INTÉRÊT.

-Si l'objet de l'enquête –la collecte de certaines variables-est une cause de non-réponse ; on devra donc les introduire parmi les variables z du calage.

Une remarque importante: quand on utilise la calage (généralisé ou pas) en situation de non-réponse, les facteurs de calage (ou g-poids) s'interprètent comme des estimations des (inverses de) probabilités de réponse.