Critère du choix des variables auxiliaires à utiliser dans l'estimateur par calage

Mohammed El Haj Tirari
Institut National de Statistique et d'Economie Appliquée - Maroc
Laboratoire de Statistique d'Enquêtes, CREST - Ensai

Septième Colloque Francophone sur les Sondages

5-7/11/2012

Introduction

- Pour estimer le total d'une population en présence d'information auxiliaire, l'estimateur par calage est parmi les plus utilisés en pratique.
  - ↪ Les poids de cet estimateur permettent de redresser l'échantillon de manière à refléter les totaux connus dans la population d'un ensemble de variables auxiliaires.
  - ↪ De plus, bien que l'estimateur par calage soit biaisé, ses poids sont calculés de telle sorte à contrôler ce biais.
- L'amélioration en termes de précision apportée par l'estimateur par calage dépend des variables auxiliaires utilisées dans le calage :
  - ↪ le biais et la variance de l'estimateur par calage sont faibles quand ces variables auxiliaires sont fortement reliées à la variable d'intérêt.

Introduction

- Cependant, la variance de l'estimateur par calage peut devenir importante quand on utilise dans le calage un très grand nombre de variables auxiliaires surtout lorsque certaines de ces variables ne sont pas reliées à la variable d'étude.

  ↪ Nécessité d'élaborer des critères permettant de sélectionner parmi ces variables celles qu'il convient d'utiliser dans le calage.

- Dans cette présentation, nous proposons un critère de sélection des variables auxiliaires qui convient d'utiliser pour calculer les poids de calage en se servant des données observées sur l'échantillon.

Notations

- Soit $U = \{1, \ldots, N\}$ une population de taille N à partir de laquelle on sélectionne un échantillon s de taille n.
- On s'intéresse à une variable d'intérêt $\mathbf{y} = (y_1, \ldots, y_N)'$ en ayant comme objectif l'estimation de son total :

$$t_y = \sum_{k \in U} y_k$$

- On suppose qu'on dispose de p variables auxilaires $X_1, \ldots, X_p$ dont les totaux

$$t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$$

sont connus, où $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})'$ pour tout $k \in U$.

Approche modèle

Sous l'approche basée sur le modèle, on suppose que les valeurs de $\mathbf{y}$ sont les réalisations d'un modèle de superpopulation ξ donné par

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \epsilon_k$$

avec

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)',$$

$$E_\xi(\epsilon_k) = 0, \quad var_\xi(\epsilon_k) = \sigma^2 v_k^2 \text{ et } cov_\xi(\epsilon_k, \epsilon_l) = 0.$$

Les v_k^2 sont supposé connus avec $\sum_{k \in U} v_k = N$

Estimateur par calage

- Pour estimer le total t_y d'une variable d'intérêt $\mathbf{y}$, on considère la classe des estimateurs linéaires qui peuvent s'écrire

$$\widehat{t}_{yw} = \sum_{k \in S} w_{kS} y_k$$

où w_{kS} sont des poids qui peuvent dépendre de l'échantillon.

- Un estimateur linéaire est dit calé sur les variables auxiliaires $\mathbf{x}_k$ si et seulement si les poids w_{kS} satisfient

$$\sum_{k \in S} w_{kS} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$$

- Le calage vise à réduire la variance des estimateurs.

Critère du choix des variables de calage

- Pour mesurer la précision de l'estimateur par calage, nous allons considèrer l'approche basée sur le plan et le modèle.

- Sous cette approche, la précision d'un estimateur linéaire est mesurée en considérant la "Variance Anticipée" définie par

$$AVar(\widehat{t}_{yw}) = E_p E_\xi (\widehat{t}_{yw} - t_y)^2$$

- La variance anticipée de l'estimateur par calage est donnée par

$$AVar(\widehat{t}_{yw}) = \sigma^2 \sum_{k \in U} v_k^2 \left[\frac{V_{kS}}{d_k} + R_{kS}^2 (d_k - 1) + (R_{kS} - 1)^2 \right]$$

où $R_{kS} = E_p(w_{kS} I_k) = \frac{E_p(w_{kS} I_k | I_k = 1)}{d_k}$
et $V_{kS} = var_p(w_{kS} \mid I_k = 1)$.

Critère du choix des variables de calage

Approximation de $AVar(\widehat{t}_{yw})$

Sous l'approche basée sur le plan et le modèle, une approximation de la précision de l'estimateur par calage peut être donnée par

$$AVar(\widehat{t}_{yw}) \approx \sigma^2 \sum_{k \in U} v_k^2 \left[R_{w_k}^2 (d_k - 1) + (R_{w_k} - 1)^2 \right]$$

avec $R_{w_k} = \frac{w_k}{d_k}$ *est le rapport des poids de calage et les poids de sondage.*

Critère du choix des variables de calage

Cette approximation a l'avantage de tenir compte des deux aspects dont dépend la précision de l'estimateur $\widehat{t}_{yw}$:

* la variance résiduelle du modèle qui diminue quand on ajoute une variable auxiliaire supplémentaire dans le modèle.

  $\hookrightarrow$ Diminution de la variance de $\widehat{t}_{yw}$.

* les rapports de poids R_{w_k} qui s'accroîent quand on ajoute une variable auxiliaire supplémentaire dans le modèle.

  $\hookrightarrow$ Augmentation du biais de $\widehat{t}_{yw}$.

Critère de choix des variables de calage

- Pour chaque variable X_j parmi les p variables auxiliaires disponibles, on peut définir

$$F_{X_j} = \frac{AVar(\widehat{t}_{yw_j})}{AVar(\widehat{t}_{yw_{j-1}})}$$

avec $\widehat{t}_{yw_{j-1}}$ est l'estimateur par calage sur les variables auxiliaires dont le pouvoir explicatif est supérieur à celui de X_j.
$\widehat{t}_{yw_j}$ est l'estimateur par calage sur la variable X_j et celles dont le pouvoir explicatif est supérieur à celui de X_j.

- F_{X_j} peut être utilisé comme un indicateur du choix ou non de la variable X_j dans le calage :

  ↬ X_j fait partie des variables de calage quand $F_{X_j} < 1$

Critère de choix des variables de calage

- F_{X_j} peut être estimé par

$$\widehat{F}_{X_j} = \frac{\widehat{AVar}(\widehat{t}_{yw_j})}{\widehat{AVar}(\widehat{t}_{yw_{j-1}})}$$

où

$$\widehat{AVar}(\widehat{t}_{yw}) = \widehat{\sigma^2} \sum_{k \in S} d_k v_k^2 \left[R_{w_k}^2 (d_k - 1) + (R_{w_k} - 1)^2 \right]$$

avec $\widehat{\sigma^2} = \frac{\|\epsilon_k\|^2}{n-p-1}$ et ϵ_k sont les résidus du modèle de régression de Y en fonction des variables auxiliaires utilisées dans le calage.

Critère de choix des variables de calage

Procédure de sélection des variables de calage

1. *La classification des variables explicatives selon leur pouvoir explicatif au moyen d'une régression de Y en fonction de toutes les variables auxiliaires disponibles.*

2. *La réalisation des calages successifs en ajoutant les variables auxiliaires une à une selon l'ordre de pouvoir explicatif de celles-ci.*

3. *A chaque étape de la sélection pas à pas, la décision de garder ou non une variable auxiliaire X_j dans le calage est basée sur le critère suivant :*

$$\widehat{F}_{X_j} = \frac{\widehat{AVar}(\widehat{t}_{yw_j})}{\widehat{AVar}(\widehat{t}_{yw_{j-1}})}$$

où on décide de garder la variable X_j dans le calage quand $\widehat{F}_{X_j} < 1$.

Remarques

- Le modèle de régression est utilisé à la première étape de la procédure pour simplifier la sélection des variables en permettant de définir l'ordre de l'inclusion de ces variables dans le calage.
- Cette procédure n'est qu'un exemple de procédures de sélection pas à pas ascendante qu'on peut considérer pour choisir les variables de calage. D'autres procédures de sélection de type ascendant peuvent être utilisées ...
- On peut également utiliser des procédures de sélection de type descendant.

Exemples

- Pour illustrer le fonctionnement de la procédure proposée du choix des variables de calage, nous avons générer un échantillon de 2006 unités sélectionnées à partir d'une population de taille 329374.

- On dispose donc des données observées sur l'échantillon pour
  - une variable d'intérêt Y,
  - des variables auxiliaires pour lesquelles on connaît le total dans la population.

- Nous avons considérer les deux modèles de régression suivants :
  1. le cas d'un modèle de régression relativement mal spécifié $(R^2 = 0,41)$,
  2. le cas d'un modèle de régression bien spécifié $(R^2 = 0,95)$.

Exemple1 : le cas d'un modèle de régression relativement mal spécifié ($R^2 = 0,41$)

	Sélection des variables selon la minimisation de la variance anticipée		Sélection des variables selon leur pouvoir explicatif dans le modèle	
Variable	$\widehat{Avar}(\hat{t}_{yw})$	$\hat{F}_{x_j}$	F Value	Pr > F
X_0	48.892000			
X_1	30.836405	0.630705	125.06	<.0001
X_2	30.116679	0.976748	4.32	<.0001
X_3	29.841137	0.991088	7.05	<.0001
X_4	30.759047	1.035218	10.51	0.0012
X_5	31.011498	1.046679	5.81	0.0031
X_6	31.717439	1.067664	5.58	0.0183
X_7	32.328059	1.085077	3.17	0.0750
X_8	32.567583	1.092551	0.51	0.4731
X_9	32.600942	1.094142	0.04	0.8341

Exemple2 : le cas d'un modèle de régression bien spécifié ($R^2 = 0,95$)

	Sélection des variables selon la minimisation de la variance anticipée		Sélection des variables selon leur pouvoir explicatif dans le modèle	
Variable	$\widehat{Avar}(\hat{t}_{w})$	$\widehat{F}_{x_j}$	F Value	Pr > F
X_0	1621.142545			
X_1	371.557951	0.229195	7519.94	<.0001
X_2	61.025967	0.164225	2037.13	<.0001
X_3	56.610056	0.927302	17.04	<.0001
X_4	55.948332	0.988979	2.31	0.0253
X_5	55.814342	0.998441	13.06	0.0003
X_6	56.679619	1.017616	0.21	0.9837
X_7	56.964903	1.023349	0.55	0.4569
X_8	56.996638	1.023878	0.37	0.5438
X_9	57.536620	1.033616	0.20	0.6572
X_{10}	57.802686	1.039157	0.17	0.6819
X_{11}	58.213983	1.047341	0.11	0.7361
X_{12}	58.345968	1.049919	0.03	0.8731
X_{13}	58.467457	1.052110	0.00	0.9883

Conclusion

- Dans ce travail, nous avons proposé un nouveau critère pour le choix des variables auxiliaires qui convient d'utiliser dans le calage.
- Ce critère se base sur l'approximation de la variance anticipée de l'estimateur par calage.
- Il a l'avantage de tenir compte du biais dû à l'utilisation des poids de calage au lieu des poids de sondage.