No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Méthodes bootstrap pour l'analyse de données d'enquête à plan de sondage complexe

J. N. K. Rao[1]

Résumé

L'application des méthodes normalisées à des données d'enquête en omettant de tenir compte des caractéristiques du plan de sondage et des redressements de la pondération peut aboutir à des inférences erronées. Les méthodes bootstrap offrent une option intéressante à l'analyste qui veut en tenir compte. Le fichier de données comprend les poids de sondage finals pour l'échantillon complet et les poids bootstrap finals connexes pour un grand nombre de répliques bootstrap, ainsi que les données observées sur les unités de l'échantillon. Nous montrons comment ce genre de fichier peut être utilisé pour analyser les données d'enquête de façon simple à l'aide d'équations d'estimation pondérées. Nous discutons aussi d'une méthode bootstrap à fonction d'estimation en une étape qui permet d'éviter certaines difficultés que pose le bootstrap.

MOTS CLÉS : poids bootstrap; caractéristiques du plan; régression logistique; équations d'estimation pondérées.

1. Introduction

Les étapes principales d'une enquête par sondage sont la conception de l'enquête, la collecte et le traitement des données, l'estimation et l'analyse des données. Dans le présent article, nous nous concentrons sur l'analyse des données provenant d'enquêtes complexes de grande portée, transversales ainsi que longitudinales. Ces enquêtes comportent souvent un échantillonnage en grappes stratifié à plusieurs degrés qui donne lieu à des dépendances entre les éléments échantillonnés, ainsi que des probabilités inégales de sélection, qui produisent des poids de sondage inégaux. De surcroît, les poids de sondage sont souvent calés sur des totaux de population connus des variables auxiliaires, et ajustés pour tenir compte de la non-réponse. Par conséquent, l'application des méthodes statistiques classiques aux données provenant de ces enquêtes sans tenir compte des caractéristiques du plan et des redressements de la pondération peut aboutir, même dans le cas de grands échantillons, à des inférences erronées, dont la sous-estimation des erreurs types des estimateurs, la surestimation des taux d'erreurs de type I (probabilité de rejeter une hypothèse nulle vraie) et des diagnostics de modélisation incorrects. Cette difficulté d'utilisation des méthodes normalisées a motivé l'élaboration de nouvelles méthodes qui tiennent compte du plan de sondage et qui donnent lieu à des inférences asymptotiquement valides. Le lecteur est prié de se référer aux ouvrages de Skinner, Holt et Smith (1989), Chambers et Skinner (2003), ainsi que Lehtonen et Pahkinen (2004) pour obtenir d'excellentes descriptions de ce genre de méthodes.

Nous nous concentrons sur les paramètres en population finie qui peuvent être formulés comme des solutions des équations d'estimation (EE) « sous recensement », ainsi que sur les paramètres d'un modèle de super-population sous-jacent qui donnent lieu aux équations d'estimation. L'approche des équations d'estimation permet de traiter une grande variété de paramètres de population finie (ou descriptifs), ainsi que de paramètres de modèle. Pour faire une inférence au sujet des paramètres d'intérêt sous échantillonnage en grappes stratifié à plusieurs degrés, nous pouvons utiliser les méthodes de linéarisation de Taylor ou les méthodes de rééchantillonnage, qui incluent le jackknife, les répliques équilibrées répétées (BRR) et le bootstrap de Rao-Wu (Rao et Wu, 1988; Rao, Wu et Yue, 1992). L'un des avantages des méthodes de rééchantillonnage est qu'une seule formule de l'erreur type est utilisée pour tous les estimateurs, contrairement à la linéarisation qui nécessite la dérivation d'une formule distincte pour chaque cas. En outre, le traitement de la poststratification et des redressements pour la non-réponse peut être

[1] J. N. K. Rao, Carleton University, School of Mathematics and Statistics, Ottawa, Canada, K1S 5B6

fastidieux dans le cas de la linéarisation, alors qu'il est relativement simple dans celui des méthodes de rééchantillonnage. Parmi ces dernières, le bootstrap offre une option intéressante à l'analyste qui veut tenir compte du plan de sondage. Il est facile à mettre en œuvre et plus souple que le jackknife et que la méthode BRR en ce qui concerne le nombre de répliques (ou échantillons répétés), et produit des inférences valides pour les statistiques lisses ainsi que non lisses (comme la médiane), contrairement au jackknife. Le fichier de données pour la mise en oeuvre du bootstrap comprend les poids finals pour l'échantillon complet et les poids bootstrap finals connexes pour un grand nombre de répliques bootstrap, ainsi que les données observées sur les éléments de l'échantillon.

Dans le présent article, nous montrons comment ce genre de fichier de données peut être utilisé couramment pour analyser des données d'enquête de manière simple à l'aide d'équations d'estimation sur échantillon pondérées (sections 2 et 3). Nous décrivons aussi une méthode bootstrap à fonction d'estimation (FE) en une étape qui permet d'éviter une difficulté que pose le bootstrap (section 4). À l'heure actuelle, Statistique Canada utilise des méthodes bootstrap pour l'analyse des données provenant de plusieurs enquêtes à grande échelle.

2. Équations d'estimation pondérées

Nous nous concentrons sur les plans de sondage stratifiés à plusieurs degrés comportant un grand nombre de strates L, et un nombre relativement faible d'unités primaires d'échantillonnage (grappes) $n_h (\geq 2)$, tirées dans chaque strate $h(=1,...,L)$. Nous supposons que le sous-échantillonnage dans les grappes échantillonnées $i(=1,...,n_h)$ est effectué de façon à s'assurer que l'estimation des totaux de grappe soit sans biais. Le poids de sondage de base attaché à l'élément hik dans l'échantillon s est dénoté par d_{hik}. Les poids d_{hik} sont calés sur les totaux de population connus des variables auxiliaires et corrigés de la non-réponse totale. Les poids finals résultant sont dénotés par $w_{hik}, hik \in s$ et présentés dans le fichier de données avec les réponses observées y_{hik} et les variables prédictives x_{hik} (éventuellement évaluées vectoriellement).

De nombreux paramètres d'intérêt peuvent être formulés en tant que solutions des équations d'estimation sous recensement

$$U(\theta) = \sum u_{hik}(\theta) = 0, \qquad (1)$$

où la sommation est faite sur les éléments de la population hik (Binder, 1983). Par exemple, les choix de $y_{hik} - \theta, I(y_{hik} \leq \theta) - \frac{1}{2}$ et $x_{hik}(y_{hik} - x'_{hik}\theta)$ pour $u_{hik}(\theta)$ dans (1) donnent la solution θ_N comme moyenne de population $\overline{Y}$, médiane de population et coefficient de régression linéaire de population, respectivement. Aucun modèle hypothétique de super-population n'est considéré et les paramètres sont descriptifs, y compris le coefficient de régression qui mesure la part de la variation de la réponse y qui est expliqué par la variable prédictive x. Cependant, le coefficient de régression peut être motivé par un modèle de régression linéaire $E_m(y_{hik}) = \mu_{hik} = x'_{hik}\theta$ avec variance de l'erreur constante, où θ est le paramètre du modèle et E_m dénote l'espérance du modèle. De même, on obtient un coefficient de régression logistique sous recensement en posant $u_{hik}(\theta) = x_{hik}(y_{hik} - \mu_{hik})$, où $\log\{\mu_{hik}/(1-\mu_{hik})\} = x'_{hik}\theta.$

Nous nous intéressons aux paramètres descriptifs (sous recensement), ainsi qu'aux paramètres du modèle θ_N lorsqu'un modèle θ est spécifié. Notons que θ_N est un estimateur convergent du paramètre du modèle θ sous un modèle spécifié. Donc, un estimateur $\hat{\theta}$ de θ_N ainsi que de θ (sous un modèle)

s'obtient en estimant le total de population $U(\theta)$ en utilisant les poids finals w_{hik} et en résolvant les équations d'estimation pondérées

$\hat{U}(\theta) = \sum_{hik \in s} w_{hik} u_{hik}(\theta) = 0$. (2)

Habituellement, on utilise l'algorithme de Newton-Raphson (NR) pour trouver la solution $\widehat{\theta}$ itérativement à partir de (2). La $(r+1)^{\text{e}}$ étape de l'itération NR est donnée par

$\hat{\theta}_{r+1} = \hat{\theta}_r + [\hat{J}(\hat{\theta}_r)]^{-1} \hat{U}(\hat{\theta}_r)$, (3)

où $\hat{\theta}_r$ est la solution à la r^{e} itération, et $\hat{J}(\hat{\theta}_r)$ et $\hat{U}(\hat{\theta}_r)$ sont les valeurs de $\hat{J}(\theta) = -\partial \hat{U}(\theta) / \partial \theta'$ et $\hat{U}(\theta)$ évaluées à $\theta = \hat{\theta}_r$. L'itération de l'algorithme jusqu'à la convergence produit la solution $\hat{\theta}$.

L'estimateur $\hat{\theta}$ est convergent sous le plan pour θ_N et convergent sous le plan et sous le modèle pour θ. En cas de non-réponse, nous devons supposer que le redressement pour la non-réponse donne lieu à des estimateurs convergents. Le lecteur trouvera dans Thompson (1997, chapitre 5) une excellente description des équations d'estimation pour les données d'enquête.

Bien que l'estimateur ponctuel soit le même pour les paramètres sous recensement et sous le modèle, les inférences diffèrent généralement. Dans le cas du paramètre sous recensement, nous estimons la variance de $\hat{\theta}$ sous le plan de sondage, tandis que dans le cas du paramètre sous le modèle, nous estimons la variance sous le plan et le modèle (ou variance totale) $V(\hat{\theta}) = E_m V_d(\hat{\theta}) + V_m E_d(\hat{\theta}) \approx E_m V_d(\hat{\theta}) + V_m(\theta_N)$, où E_d, V_d dénotent l'espérance sous le plan et la variance sous le plan, et V_m dénote la variance sous le modèle. Un estimateur robuste du premier terme $E_m V_d(\hat{\theta})$ s'obtient en estimant la variance sous le plan $V_d(\hat{\theta})$, mais l'estimation du deuxième terme $V_m(\theta_N)$ requiert la spécification de la structure de covariance sous le modèle. Cependant, pour les plans d'échantillonnage en grappes à plusieurs degrés, le deuxième terme est petit comparativement au premier si n/N est négligeable, où n et N dénotent le nombre total de grappes échantillonnées et le nombre total de grappes dans la population. Dans ce cas, l'estimateur de la variance sous le plan peut aussi être utilisé pour estimer la variance totale.

3. Bootstrap de Rao-Wu

Dans la méthode du bootstrap de Rao-Wu, un échantillon bootstrap est obtenu en tirant un échantillon aléatoire simple de $m_h = n_h - 1$ grappes avec remise à partir des n_h grappes échantillonnées, indépendamment pour chaque strate h.

Nous sélectionnons un grand nombre, B, d'échantillons bootstrap indépendants qui peuvent être représentés en fonction des fréquences bootstrap $m_{hi}(b) =$ nombre de fois que la $(hi)^{\text{e}}$ grappe d'échantillon est sélectionnée dans le b^{e} échantillon bootstrap ($b = 1,...,B$). Habituellement, $B = 500$ échantillons bootstrap sont utilisés à Statistique Canada. Les poids de sondage bootstrap sont simplement donnés par $d_{hik}(b) = d_{hik}\{n_h/(n_h - 1) m_{hi}(b)\}$. Maintenant, nous remplaçons d_{hik} par $d_{hik}(b)$ dans la méthode d'obtention des poids finals à partir de d_{hik} pour obtenir les poids bootstrap final $w_{hik}(b)$. Le fichier de données des poids sera constitué de $\{w_{hik}, w_{hik}(1),...w_{hik}(B)\}$. Le lecteur est prié de se référer à Girard (2006) pour les questions pratiques relatives à la génération de poids bootstrap dans le contexte de l'Enquête longitudinale nationale sur les enfants et les jeunes réalisée au Canada.

La substitution des poids bootstrap $w_{hik}(b)$ à w_{hik} dans les équations d'estimation pondérées (2) nous donne les équations d'estimation bootstrap

$$\hat{U}_b(\theta) = \sum_{hik \in s} w_{hik}(b) u_{hik}(\theta) = 0 \qquad (4)$$

Il suffit d'exécuter une itération NR en une étape de (4) en prenant l'estimation sur échantillon complet $\hat{\theta}$ comme valeur de départ pour obtenir les estimations bootstrap $\hat{\theta}(b)$ données par

$$\hat{\theta}(b) = \hat{\theta} + [\hat{J}_b(\hat{\theta})]^{-1} \hat{U}_b(\hat{\theta}), \qquad (5)$$

où $\hat{J}_b(\theta) = -\partial \hat{U}_b(\theta) / \partial \theta'$, à condition que $\hat{J}_b(\hat{\theta})$ soit inversible. Cependant, Binder et coll. (2004) ont trouvé qu'il était possible d'avoir plusieurs échantillons bootstrap b pour lesquels $\hat{J}_b(\hat{\theta})$ est incorrectement conditionnée et donc non inversible.

En supposant que $\hat{\theta}(b)$ est disponible pour tous les échantillons bootstrap, l'estimateur bootstrap de Rao-Wu de la matrice de covariance de $\hat{\theta}$ (dénoté estimateur de la variance de $\hat{\theta}$) est donné par

$$v_B(\hat{\theta}) = B^{-1} \sum_{b=1}^{B} [\hat{\theta}(b) - \hat{\theta}][\hat{\theta}(b) - \hat{\theta}]' \qquad (6).$$

La racine carrée des éléments diagonaux de (6) donnera les erreurs types bootstrap des composantes du vecteur $\hat{\theta}$. Un autre estimateur de la variance par le bootstrap est obtenu en substituant $\hat{\theta}(.) = B^{-1} \sum_b \hat{\theta}(b)$ à $\hat{\theta}$ dans (6). Les deux estimateurs bootstrap de la variance sont asymptotiquement équivalents, mais les erreurs types obtenues d'après (6) seront un peu plus grandes que celles provenant de l'estimateur bootstrap de la variance de rechange. Les deux estimateurs bootstrap de la variance aboutissent à une surestimation de la variance sous le plan $V_d(\hat{\theta})$, mais cette surestimation sera faible si la fraction d'échantillonnage globale de premier degré n/N est faible, même si les fractions d'échantillonnage de premier degré n_h / N_h ne sont pas négligeables pour certaines strates h (Shao, 2002). En outre, la surestimation de la variance sous le plan pourrait en fait être utile lors de l'estimation de la variance totale de $\hat{\theta}$ et donc de l'inférence sur les paramètres du modèle θ.

En utilisant l'estimateur $\hat{\theta}$ et l'estimateur bootstrap de la variance connexe, on peut également obtenir des intervalles de confiance en grand échantillon pour les composantes de θ et effectuer des tests de Wald et du quai-score (Rao, Scott et Skinner, 1998, Rao,1999) pour vérifier les hypothèses concernant θ. Il est supposé que le nombre de grappes dans l'échantillon de premier degré, n, est grand.

Beaumont et Bocci (2006) ont appliqué la méthode du bootstrap pour trouver des valeurs P pour tester une hypothèse linéaire de la forme $H_0 : A\theta = c$ sous un modèle de régression linéaire $y_{hik} = x'_{hik}\theta + \varepsilon_{hik}$, où A est une matrice de plein rang spécifiée, c est un vecteur spécifié et ε_{hik} représente les erreurs de modélisation. Une statistique F pondérée par les poids de sondage, $F(c)$, pour tester H_0 peut être obtenue en utilisant l'énoncé WEIGHT dans la procédure REG de SAS, à condition que tous les poids d'échantillonnage finals soient positifs. Le traitement de cette statistique comme une variable F donne lieu à des valeurs P incorrectes, même pour les grands échantillons. Les poids bootstrap peuvent être utilisés pour trouver des valeurs P valides. Nous utilisons simplement l'énoncé WEIGHT de SAS avec w_{hik} remplacé par les poids bootstrap $w_{hik}(b)$ et c dans H_0 remplacé

par $A\hat{\theta}$ pour obtenir la statistique F pondérée par les poids bootstrap, $F_b(A\hat{\theta})$, pour chaque échantillon bootstrap $b = 1,...,B$. La valeur P bootstrap est alors donnée par la proportion de valeurs F bootstrap, $F_b(A\hat{\theta})$, supérieure à la valeur F en échantillon complet, $F(c)$. L'hypothèse nulle est rejetée si la valeur P bootstrap est inférieure à un seuil prescrit, disons 0,05. Une étude par simulation donne à penser que la méthode proposée a de bonnes propriétés de puissance et qu'elle maintient la taille du test. Il serait utile de l'étendre à d'autres modèles et à d'autres hypothèses. Notons que la méthode est conçue pour la vérification d'hypothèses.

4. Bootstrap de la fonction d'estimation

À la section 3, nous avons noté que la méthode directe du bootstrap de Rao-Wu peut poser des difficultés si les matrices sont mal conditionnées lors de l'exécution des itérations NR pour obtenir les estimations bootstrap de θ. Pour contourner ce problème, Rao et Tausi (2004) ont proposé une approche par bootstrap de la fonction d'estimation (FE), motivée par les travaux de Hu et Kalbfleisch (2000) dans le cas de données ne provenant pas d'un sondage. Nous approximons la fonction d'estimation $\hat{U}(\theta)$ par la fonction d'estimation bootstrap $\hat{U}_b(\hat{\theta})$ et résolvons

$$\hat{U}(\theta) = \hat{U}_b(\hat{\theta}) \qquad (7)$$

pour trouver θ par une itération NR en une étape en prenant $\hat{\theta}$ comme valeur de départ pour obtenir

$$\tilde{\theta}(b) = \hat{\theta} - [\hat{J}(\hat{\theta})]^{-1}\hat{U}_b(\hat{\theta}). \qquad (8)$$

Notons que, pour tout $b = 1,...,B$, la matrice inverse dans (8) est basée sur l'échantillon complet et demeure donc la même, de sorte qu'il n'est pas nécessaire de répéter l'inversion pour chaque échantillon bootstrap. Cela évite donc le problème des matrices mal conditionnées de la méthode directe du bootstrap et simplifie également les calculs.

Pour obtenir l'estimateur de la variance de $\hat{\theta}$ par bootstrap de la FE, nous remplaçons simplement $\hat{\theta}(b)$ dans (6) par :

$$v_{EFB}(\hat{\theta}) = B^{-1}\sum_{b=1}^{B}[\tilde{\theta}(b) - \hat{\theta}][\tilde{\theta}(b) - \hat{\theta}]'. \qquad (9)$$

Binder, Kovacevic et Roberts (2004) ont approximé $\hat{U}(\theta)$ by $-\hat{U}_b(\hat{\theta})$ au lieu de $\hat{U}_b(\hat{\theta})$, mais l'estimateur de la variance par bootstrap de la FE résultant (6) basé sur l'itération NR en une étape est algébriquement identique à l'estimateur de la variance par bootstrap de Rao-Tausi. Hu et Kalbfleisch (2000) ont montré que l'approximation de $\hat{U}(\theta)$ par $\hat{U}_b(\hat{\theta})$ donne des intervalles de confiance ayant de bonnes propriétés.

Roberts, Rao et Ren (2006) ont considéré des modèles de régression logistique marginaux avec des données sur des réponses binaires provenant d'une enquête longitudinale à plan de sondage complexe. Ils ont adapté la méthode des équations d'estimation généralisées (EEG) de Liang et Zeger (1986) au cas des données d'enquête longitudinales et ont estimé les paramètres du modèle en tant que solutions des EEG pondérées par les poids de sondage. Ils ont utilisé l'approche des rapports de cotes pour modéliser la structure de covariance de travail pour les mesures répétées sur les individus échantillonnés. Les inférences sur les paramètres du modèle étaient fondées sur les erreurs types par bootstrap de la FE qui tiennent compte de la dépendance intra-sujet des mesures répétées, ainsi que de la mise en grappes des sujets et d'autres caractéristiques du plan de sondage. Ils ont appliqué la théorie aux données longitudinales provenant de l'Enquête nationale sur la santé de la population (ENSP) réalisée par Statistique Canada. L'ENSP a débuté en 1994-1995 et recueille des renseignements tous les deux ans auprès des mêmes

personnes échantillonnées. Un plan d'échantillonnage en grappes stratifié à plusieurs degrés a été utilisé pour sélectionner les ménages dans les grappes, puis pour sélectionner dans chaque ménage échantillonné un membre de 12 ans et plus pour faire partie du panel longitudinal.

Références

Beaumont, J.-F., et Bocci, C. (2006). "A Practical Bootstrap Method for Testing Hypotheses from Survey Data", unpublished report, Ottawa, Canada: Statistics Canada.

Binder, D. (1983), "On the Variance of Asymptotically Normal Estimators from Complex Surveys", *International Statistical Review*, 51, pp. 279-293.

Binder, D., Kovacevic, M, et Roberts, G. (2004), "Design-based Methods for Survey Data: Alternative Uses of Estimating Functions", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp.

Chambers, R., et Skinner, C. J. (2003), *Analysis of Survey Data*, Chichester: Wiley.

Girard, C. (2006), "How to Avoid Getting All Tied Up Bootstrapping a Survey: A Walk-Through Featuring the National Longitdinal Survey of Children and Youth", unpublished report (draft version), Ottawa, Canada: Statistics Canada.

Hu, F., et Kalbfleisch, J.D. (2000), "The Estimating Function Bootstrap", *Canadian Journal of Statistics*, 28, pp. 449-499.

Lehtonen, R., et Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys, 2nd Edition*, Chichester: New York.

Liang, K.-Y., et Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models". *Biometrika*, 73, pp. 13-22.

Rao, J. N. K., et Wu, C. F. J. (1988), "Resampling Inference with Complex Survey Data", *Journal of the American Statistical Association*, 83, pp. 231-241.

Rao, J. N. K., Wu, C. F. J., et Yue, K. (1992), "Some Recent Work on Resampling Methods", *Survey Methodology*, 18, pp. 209-217.

Rao, J. N. K., Scott, A. J., et Skinner, C. J. (1998), "Quasi-Score Tests with Survey Data", *Statistica Sinica*, 8, 1059-1070.

Rao, J. N.K. (1999), "Some Current Trends in Sample Survey Theory and Methods", *Sankhya B,* 61, pp. 1-57.

Rao, J. N. K., et Tausi, M. (2004), "Estimating Function Jackknife Variance Estimators Under Stratified Multistage Sampling", *Communications in Statistics*, 33, pp. 2087-2095.

Roberts, G., Ren, Q., et Rao, J. N. K. (2006), "Using Marginal Mean Models with Data from a Longitudinal Survey Having a Complex Design: Some Advances in Methods", to appear in *Methodology of Longitudinal Surveys*, Chichester: Wiley.

Shao, J. (2002), "Resampling Methods for Variance Estimation in Complex Surveys with Imputed Data", in R. M. Groves et al. (eds) *Survey Nonresponse*, New York: Wiley, pp. 303-314.

Skinner, C. J., Holt, D., et Smith, T. M. F. (1989), *Analysis of Complex Surveys*, New York: Wiley.

Thompson, M. E. (1997), *Theory of Sample Surveys*, London: Chapman and Hall.