Propriétés asymptotiques d'estimateurs non paramétriques model-based de la fonction de répartition sur un petit domaine

Sandrine Casanova, Eve Leconte

Propriétés asymptotiques d'estimateurs non paramétriques model-based de la fonction de répartition sur un petit domaine

Sandrine Casanova & Eve Leconte

TSE (GREMAQ), Université Toulouse 1 Capitole,
21, allée de Brienne, 31000 TOULOUSE, France

E-mail : sandrine.casanova@TSE-fr.eu, eve.leconte@TSE-fr.eu

Résumé

Nous nous intéressons à l'estimation de la fonction de répartition (f.d.r.) en sondage sur des sous-populations (domaines). Si un domaine est de taille suffisante, l'estimation de la f.d.r. se base uniquement sur les individus du domaine et les estimateurs produits sont de précision acceptable. Cependant, dans la plupart des applications, les tailles d'échantillons correspondant à des petits domaines ne sont pas suffisantes. L'estimation se fonde alors sur une information auxiliaire fournie par une covariable et de l'information est "empruntée" aux autres domaines. Dans ce contexte, Chambers et Tzavidis (2006) ont proposé un estimateur paramétrique de la f.d.r. sur un domaine, basé sur des quantiles. Casanova (2007, 2010) a adapté cet estimateur au cas non paramétrique et a proposé un autre estimateur basé sur les quantiles. Ces estimateurs se placent dans un cadre *model-based* où le problème est de prédire la variable d'intérêt pour les individus non échantillonnés du domaine. Pour un individu fixé, sa variable d'intérêt peut toujours être vue comme le quantile conditionnel à la valeur de sa covariable pour un certain ordre appelé ordre-quantile. Les ordres-quantiles des individus de l'ensemble des échantillons sont estimés et on prédit ensuite à l'aide de polynômes locaux la variable d'intérêt d'un individu hors échantillon par les quantiles conditionnels associés aux ordres qui décrivent ou résument le domaine de l'individu. Nous nous focalisons ici sur les propriétés asymptotiques de ces estimateurs avec une approche *model-based* : nous étudions leur biais asymptotique sous le modèle ainsi que leur convergence en moyenne quadratique.

Abstract

We work on estimating the cumulative distribution function (c.d.f.) in survey sampling on a sub-population (domain). If the size of the domain is large enough, the estimation of the c.d.f. relies on data from sample units in the domain and the resultant estimates will be of acceptable precision. However, in most practical applications, sample sizes are not large enough to produce sufficiently precise estimators. In such situations, the estimation

is based on auxiliary information related to the variable of interest and information is "borrowed" from the other domains. In this framework, Chambers et Tzavidis (2006) have proposed a parametric estimator of the c.d.f. in a domain based on quantiles. Casanova (2007, 2010) has adapted this estimator to a nonparametric approach and has proposed another estimator based on quantiles. The considered estimators are in a model-based framework where the problem is the prediction of the interest variable for the non-sampled units of the domain. For a given unit, the interest variable can be seen as the conditional quantile to the covariate of the unit for an order called quantile-order. The quantile-orders of the sample observations are estimated and local polynomials techniques are then used to predict the interest variable for the non-sampled individuals of the domain by conditional quantiles associated to the quantile-orders which describe or resume the domain. We focus here on asymptotic properties of the new estimators in a model-based framework : asymptotic bias under the model and mean squared error convergence will be considered.

Mots-clés : sondages, fonction de répartition, information auxiliaire, model-based, domaine, polynômes locaux, quantiles conditionnels, propriétés asymptotiques.

1 Introduction

Soit une population U partitionnée en m sous-populations ou domaines U_i de taille N_i, $i = 1, \cdots, m$. Soient s un échantillon de U de taille n et $s_i = s \cap U_i$ un échantillon du domaine U_i de taille n_i. En sondage, on s'intéresse à l'estimation de la fonction de répartition (f.d.r.) d'une variable d'intérêt Y sur le domaine U_i définie par : $F_i(y) = \frac{1}{N_i} \sum_{j \in U_i} \mathbb{1}(y_{ij} \leq y)$ que l'on peut décomposer en

$$F_i(y) = \frac{1}{N_i} \left(\sum_{i \in s_j} \mathbb{1}(y_{ij} \leq y) + \sum_{j \in U_i \setminus s_i} \mathbb{1}(y_{ij} \leq y) \right) \tag{1}$$

où y_{ij} est la variable d'intérêt mesurée pour le j-ième individu du domaine U_i . On suppose que y_{ij} est seulement connue sur s_i.

Un estimateur naturel de la f.d.r. sur le domaine U_i est la fonction de répartition empirique définie par :

$$\hat{F}_i^{Emp}(y) = \frac{1}{n_i} \sum_{j \in s_i} \mathbb{1}(y_{ij} \leq y).$$

Si la taille d'échantillon est trop faible, ce qui est le cas pour de petits domaines, on peut améliorer l'estimation à l'aide d'une information auxiliaire apportée par une covariable et "emprunter de la force aux voisins". Ceci nous permet d'estimer le deuxième terme dans la décomposition de la f.d.r. (1) en estimant les y_{ij} des individus non échantillonnés.

Dans ce cadre, Casanova (2007, 2010) a proposé deux estimateurs non paramétriques de la f.d.r sur un domaine en prédisant les y_{ij} des individus non échantillonnés par des quantiles conditionnels estimés à l'aide des polynômes locaux. Le deuxième estimateur est une adaptation d'un estimateur proposé dans un cadre paramétrique par Chambers et Tzavidis (2006). Les deux méthodes d'estimation sont rappelées dans la section 2. En section 3, nous nous intéressons aux propriétés asymptotiques de ces estimateurs.

2 Estimation non paramétrique de la f.d.r. sur un petit domaine

Les deux estimateurs proposés par Casanova (2007, 2010) comportent deux étapes, la première étant commune aux deux estimateurs.

Etape 1 : estimation non paramétrique des ordres-quantiles des points échantillonnés de l'ensemble des domaines

Pour une observation (y_k, x_k), il existe un ordre $q_k \in [0,1]$ tel que y_k est le quantile conditionnel à x_k d'ordre q_k. Cet ordre, appelé ordre-quantile conditionnel, classe l'observation dans l'échantillon. Les ordres-quantiles conditionnels des observations d'un domaine le situent par rapport à l'ensemble de tous les domaines. Une estimation naturelle de l'ordre-quantile conditionnel peut se faire à l'aide de l'estimateur de Nadaraya-Watson de la f.d.r. conditionnelle :

$\hat{q}_k(y_k, x_k) = \dfrac{\sum_{l \in s} \mathbb{1}(y_l \leq y_k) K\left(\dfrac{x_k - x_l}{h}\right)}{\sum_{l \in s} K\left(\dfrac{x_k - x_l}{h}\right)}$, où K est un noyau de densité et h une fenêtre appropriée.

Chaque domaine U_i peut donc être décrit par l'ensemble des ordres-quantiles estimés sur le sous-échantillon s_i. Nous noterons ces ordres $\{\hat{q}_{ik},\ k = 1, \dots n_i\}$. Alternativement, la position du domaine U_i dans la population peut être résumée par l'ordre-quantile moyen des points échantillonnés du domaine, noté $\hat{q}_i = \dfrac{1}{n_i} \sum_{k \in s_i} \hat{q}_{ik}$.

Etape 2 : estimations de la variable d'intérêt pour les points non échantillonnés du domaine

Pour incorporer l'information auxiliaire apportée par la variable x, on suppose le modèle de superpopulation suivant, noté ξ :

$$y_{ij} = m(q, x_{ij}) + \varepsilon_{ij}$$

où q est un réel fixé de $[0,1]$ et où les ε_{ij} sont i.i.d. de distribution G tels que $P(\varepsilon_{ij} \leq 0 \mid X_{ij} = x_{ij}) = q$. Sous ce modèle, $m(q, x_{ij})$ est le quantile de y_{ij} conditionnnellement à x_{ij}.

• **Premier estimateur :**

Pour chaque individu j de $U_i \setminus s_i$ de covariable x_{ij}, comme son domaine est décrit par les n_i ordres-quantiles $\{\hat{q}_{ik},\ k = 1, \dots n_i\}$, on peut donc calculer n_i prédictions $\hat{m}(\hat{q}_{ik}, x_{ij})$ de y_{ij} à l'aide d'estimateurs des quantiles conditionnels à x_{ij} d'ordre $\hat{q}_{ik}$.

Casanova (2007, 2010) propose d'utiliser les polynômes locaux (constante linéaire) pour estimer ces quantiles conditionnels selon le modèle ξ en utilisant les individus échantillonnés de tous les domaines. La méthode de la constante linéaire pour estimer un quantile conditionnel consiste, pour un x_{ij} donné et un ordre $\hat{q}_{ik}$ fixé, à trouver θ minimisant :

$$\sum_{l \in s} \rho_{\hat{q}_{ik}}(y_l - \theta) K(\frac{x_{ij} - x_l}{h}) \tag{2}$$

où ρ_q est la fonction de perte suivante :

$$\rho_q(u) = \begin{cases} -(1-q)u & \text{si} \quad u < 0 \\ qu & \text{si} \quad u \geq 0. \end{cases}$$

Un estimateur de la f.d.r. sur le domaine U_i s'en déduit donc d'après la formule (1) :

$$\hat{F}_i^C(y) = \frac{1}{N_i} \left(\sum_{j \in s_i} \mathbb{1}(y_{ij} \leq y) + \sum_{j \in U_i \setminus s_i} \left(\frac{1}{n_i} \sum_{k \in s_i} \mathbb{1}(\hat{m}(\hat{q}_{ik}, x_{ij}) \leq y) \right) \right)$$

Cette estimation utilise l'échantillon s de la population totale et de la force est donc empruntée à tous les domaines.

• **Deuxième estimateur :**

Cet estimateur adapte au cas non paramétrique la technique de Chambers et Tzavidis (2006), qui considère que chaque domaine U_i peut être résumé par son ordre-quantile moyen $\hat{q}_i$. Pour chaque domaine U_i fixé, on estime à l'aide de l'échantillon total s, par la méthode de la constante linéaire, les n_i quantiles conditionnels $\{m(\hat{q}_i, x_{ik}),\ k \in s_i\}$ par la formule (2). Cela permet de calculer les n_i résidus des points échantillonnés de ce domaine :

$$\hat{\epsilon}_{ik} = y_{ik} - \hat{m}(\hat{q}_i, x_{ik}), \quad k \in s_i$$

.

Pour chaque individu non échantillonné de U_i de covariable x_{ij}, ces n_i résidus pemettent de construire n_i prédictions de la forme $\hat{m}(\hat{q}_i, x_{ij}) + \hat{\epsilon}_{ik}$ de y_{ij}, où les $\hat{m}(\hat{q}_i, x_{ij})$ sont encore calculés grâce à la formule (2). L'estimateur de la f.d.r. de Y dans le domaine U_i qui en résulte est alors défini par :

$$\hat{F}_i^{CT}(y) = \frac{1}{N_i} \left(\sum_{j \in s_i} \mathbb{1}(y_{ij} \leq y) + \sum_{j \in U_i \setminus s_i} \left(\frac{1}{n_i} \sum_{k \in s_i} \mathbb{1}(\hat{m}(\hat{q}_i, x_{ij}) + \hat{\epsilon}_{ik} \leq y) \right) \right).$$

3 Propriétés asymptotiques

Casanova (2010) a montré à l'aide d'études de simulation l'apport de ces méthodes par rapport à l'estimateur empirique de la fonction de répartition sur le domaine.

Nous allons nous intéresser dans cette présentation à deux propriétés asymptotiques de ces nouveaux estimateurs. Dans un premier temps, nous allons étudier leur biais asymptotique sous le modèle. Ensuite, nous envisagerons la convergence de ces estimateurs en moyenne quadratique. Pour cela, nous nous inspirerons des techniques utilisées dans les articles de Chambers *et al.* (1992), Dorfman et Hall (1993) et Johnson *et al.* (2004), auteurs qui se sont intéressés à la convergence d'estimateurs *model-based* de la f.d.r. en population finie.

4 Bibliographie

[1] Aragon Y. et Casanova S. (2007). Estimation de la fonction de répartition sur un domaine à l'aide des quantiles et des M-quantiles conditionnels,XXXIXèmes Journées de Statistique, Angers.
[2] Casanova S. (2010). Using M-quantiles to estimate a cumulative distribution function in a domain. En révision aux *Annales d'Economie et de Statistique*.
[3] Chambers, R. L., Dorfman, A. H. and Hall, P. (1992). Properties of estimators of the finite population distribution function. Biometrika, 79, 3, 577-582.
[4] Chambers, R. L. and Tzavidis, N. (2006). M-quantiles models for small area estimation, Biometrika, 93, 255–268.
[5] Dorfman, A.H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21, 3, 1452–1475.
[6] Johnson, A., A., Breidt, F. J. and Opsomer, J. D. (2008), Estimating distribution functions from survey data using nonparametric regression. *Journal of Statistical Theory and Practice*, 2, 419–431.