Journées de méthodologie statistique de l'Insee 2022

Une approche multi-robuste dans le cadre de l'intégration de données

Sixia Chen (*), David Haziza (**)

(*) Hudson College of Public Health, University of Oklahoma
(**) Department of mathematics and statistics, University of Ottawa

dhaziza@uottawa.ca

Mots-clés. : Équations estimantes, Estimation par score de propension, Estimation de la variance, Imputation fractionelle, Modèle de participation.

Domaines. Intégration de données.

Résumé

Traditionnellement, les instituts nationaux de statistique (INS) ont collecté les données au moyen de procédures d'échantillonnage probabiliste et les inférences ont été effectuées par rapport au plan de sondage. Dans le cas d'une inférence basée sur le plan de sondage, les propriétés des estimateurs ponctuels et de variance sont évaluées par rapport au plan de sondage et des inférences valides peuvent être tirées sans s'appuyer sur la validité d'un modèle, à condition que les erreurs non dues à l'échantillonnage soient négligeables. Cela ne veut pas dire que les modèles ne peuvent pas jouer un rôle dans le cadre inférentiel basé sur le plan de sondage. En effet, l'efficacité des estimateurs ponctuels peut être améliorée en utilisant une information auxiliaire au stade de l'estimation, en capitalisant sur la relation entre une variable d'intérêt et un ensemble de prédicteurs. Les procédures d'estimation qui en résultent, appelées procédures assistées par un modèle, utilisent un modèle de travail comme véhicule afin de construire des estimateurs ponctuels, mais les estimateurs qui en résultent restent convergents par rapport plan de sondage même si le modèle est mal spécifié ; voir, par exemple, Särndal et col. (1992) et Bredit et Opsomer (2017).

Ces dernières années, les méthodes d'intégration des données provenant d'échantillons probabilistes et non probabilistes ont fait l'objet d'une grande attention, car la diminution des taux de réponse et l'augmentation des coûts de collecte des données sont devenues une préoccupation majeure. De nos jours, divers types de sources de données non probabilistes sont à la disposition des praticiens d'enquêtes, notamment les panels opt-in, les médias sociaux et les informations satellitaires. Bien que ces sources de données fournissent des données actuelles pour un grand nombre de variables et d'unités de la population, elles ne parviennent souvent pas à représenter la population cible d'intérêt en raison de biais de sélection inhérents. La présence (ou la participation) d'une unité sur une source non probabiliste est inconnue, contrairement aux procédures

d'échantillonnage probabiliste, pour lesquelles les probabilités d'inclusion sont connues, en général. La manière d'intégrer des données provenant d'échantillons non probabilistes a suscité beaucoup d'attention ces dernières années. Le lecteur est invité à consulter les articles suivants : Rivers (2007), Bethlehem (2016), Elliot et Vaillant (2017), Lohr et Raghunathan (2017), Chen et aol. (2019), Beaumont (2020), et Rao (2020) pour des discussions sur les méthodes d'intégration de données.

Les procédures d'estimation de l'intégration des données peuvent être classées en trois grandes catégories : (i) pondération par calage d'un échantillon non probabiliste sur des totaux estimés à partir d'une enquête probabiliste, par exemple, Elliot et Vaillant (2017) ; (ii) appariement statistique ou imputation de masse, par exemple, Rivers (2007) ; et (iii) pondération par score de propension d'un échantillon non probabiliste, par exemple, Chen et col. (2019). Cependant, quelle que soit l'approche utilisée, la validité des estimateurs ponctuels dépend fortement de la validité du modèle supposé. Par conséquent, les estimateurs ponctuels sont vulnérables à une mauvaise spécification du modèle. Afin de fournir une certaine robustesse contre la mauvaise spécification du modèle, Chen et col. (2019) ont proposé un estimateur doublement robuste de la moyenne dUne population, incorporant les prédictions obtenues par l'ajustement d'un modèle dd'imputation et les estimations des probabilités de participation obtenues par l'ajustement d'un modèle de participation. Les procédures doublement robustes sont intéressantes car elles offrent une certaine protection contre la mauvaise spécification de l'un ou l'autre modèle. Cependant, les procédures doublement robustes ont tendance à exhiber des performances numériques médiocres lorsque les deux modèles sont mal spécifiés.

Les principales contributions de ce projet sont les suivantes : (i) nous proposons un cadre inférentiel général multi-robuste dans le contexte de l'intégration de données. Dans ce contexte, cet article constitue, à notre connaissance, la première tentative de développer des procédures d'imputation de masse et d'estimation par score de propension qui peuvent être basées sur de multiples modèles d'imputation et/ou de multiples modèles de participation. Chaque modèle peut être basé sur des fonctionnels différentes et/ou des ensembles de variables explicatives différents. Les estimateurs qui en résultent sont dits multi-robustes au sens où ils restent convergents si tous les modèles sauf un sont mal spécifiés. Dans le contexte des données manquantes, les procédures multi-robustes ont été étudiées par Han et Wang (2013), Chan et Yam (2014), Han (2014a), Han (2014b), Chen et Haziza (2017), Duan et Yin (2017) et Chen et Haziza (2019). Un certain nombre d'études empiriques ont suggéré que, contrairement aux procédures doublement robustes, les procédures multi-robustes ont tendance à exhiber de bonnes performances numériques même lorsque tous les modèles sont mal spécifiés, ce qui est une caractéristique intéressante ; voir Han (2014a) et Chen et Haziza (2107). L'incorporation de multiples modèles de participation est particulièrement intéressante dans le contexte de l'intégration de données car, contrairement au cas des données manquantes où les variables explicatives sont observées à la fois pour les répondants et les non-répondants, ces dernières ne sont généralement pas observées pour les unités qui n'ont pas participé à l'échantillon non probabiliste. (ii) Les méthodes proposées peuvent s'appliquer à tout paramètre qui peut être exprimé comme solution d'une équation estimante. Cela comprend les moyennes/totaux de population, les fonctions de répartition et les quantiles. (iii) Des procédures multi-robustes pour les quantiles ont été proposées par Han et col. (2019) dans un contexte de données manquantes. Cependant, leurs procédures étaient basées sur des modèles de régression des résultats complètement paramétriques. Nos méthodes sont basées sur des modèles d'imputation semi-paramétriques qui ne nécessitent pas d'hypothèses de distribution sur les erreurs du modèle, contrairement à Han et al. (2019).

Bibliographie

[1] Beaumont, J. (2020). Are probability surveys bound to disappear for the production of official statistics ?, *Survey Methodology* 46(1), 1-28.

[2] Bethlehem, J. (2016). Solving the nonresponse problem with sample matching ?, *Social Science Computer Review* 34(1), 59-77.

[3] Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques, *Statistical Science* 32(2), 190-205.

[4] Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem, *Statistical Science* 29(3), 380-396.

[5] Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys, *Biometrika* 104(2), 439-453.

[6] Chen, S. and Haziza, D. (2019). Multiply robust nonparametric multiple imputation for the treatment of missing data, *Statistica Sinica* 29(4), 2035-2053.

[7] Chen, S. and Kim, J. K. (2017). Semiparametric fractional imputation using empirical likelihood in survey sampling, *Statistical theory and related fields* 1(1), 69-81.

[8] Chen, Y., Li, P. and Wu, C. (2019). Doubly robust infer- ence with nonprobability survey samples, Journal of the American Statistical Association pp. DOI : 10.1080/01621459.2019.1677241, https ://www.tandfonline.com/doi/abs/10.1080/01621459.2019.1677241.

[9] Duan, X. and Yin, G. (2017). Ensemble approaches to estimating the population mean with missing response, *Scandinavian Journal of Statistics* 44(4), 899-917.

[10] Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples, *Statistical Science* 32(2), 249-264.

[11] Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis, *Journal of Statistical Planning and Inference* 148, 101-110.

[12] Han, P. (2014b). Multiply robust estimation in regression analysis with miss- ing data, *Journal of American Statistical Association* 109(507), 1159-1173.

[13] Han, P., Kong, L., Zhao, J. and Zhou, X. (2019). A general framework for quantile estimation with incomplete data, *Journal of the Royal Statistical Society : Series B* (Statistical Methodology) 81(2), 305-333.

[14] Han, P. and Wang, L. (2013). Estimation with missing data : beyond double robustness, *Biometrika* 100(2), 417-430.

[15] Lohr, S. L. and Raghunathan, T. E. (2017). Combining survey data with other data sources, *Statistical Science* 32(2), 293-312.

[16] Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources, *Sankhya B*, In Process.

[17] Rivers, D. (2007). Sampling for web surveys, *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

[18] Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling.* Springer.