AGRÉGATION DE DONNÉES MULTIMODE : IMPACT SUR LA MODÉLISATION DES VARIABLES PRÉSENTANT UN EFFET DE MESURE

Christophe BARRET, Mady CISSÉ

Céreq, Centre d'Études et de REcherches sur les Qualifications

mady.cisse@cereq.fr

Mots-clés : Multimode, agrégation, imputation, matching, estimation.

Résumé

Les enquêtes Génération du Céreq permettent d'étudier les parcours d'insertion professionnelle des jeunes à la sortie du système éducatif, et font l'objet d'un processus de rénovation dans lequel une collecte multimode Internet/Téléphone est envisagée. Des expérimentations de l'utilisation d'Internet comme mode de collecte ont donc été menées en parallèle des enquêtes par téléphone en 2013, 2015 et 2016.

La combinaison de plusieurs modes de collecte plutôt qu'un seul implique deux effets qui modifient l'estimation des paramètres d'intérêt. D'une part il y a des effets de mesure : les enquêtés ne répondent pas nécessairement de la même manière selon le mode de collecte. D'autre part il y a l'effet de sélection : selon le mode de collecte on va atteindre principalement les enquêtés qui y sont le plus réceptifs.

Une méthode de matching sur score de propension a été employée pour distinguer les effets de sélection des effets de mesure, et ainsi repérer les variables sujettes aux effets de mesure. Ces derniers sont problématiques lors de la réalisation d'estimations car, en l'absence de traitements spécifiques, ils sont sources de biais. Dans le cadre, d'enquêtes répétées, et toujours en l'absence de traitements spécifiques, ces biais sont sensibles à la variation de la répartition des modes de collecte entre les différentes enquêtes.

Cette présentation fait suite à celle présentée lors des dernières JMS [1] qui présentait la détection de ces effets au sein des enquêtes Génération ainsi que des éléments sur le protocole de collecte multimode. Il s'agira ici de comparer différentes approches d'agrégation des données, dont certaines tentent de corriger le biais lié à l'effet de mesure.

La première, l'agrégation simple, consiste à juxtaposer les données collectées dans chacun des deux modes sans appliquer de traitement spécifique au mode sur les réponses.

La seconde est une méthode proposée à titre expérimental par l'INSEE[3] qui repose sur l'imputation des réponses des individus qui portent l'effet de mesure sur les variables d'intérêt. Les individus ayant des réponses, sur les variables présentant un effet de mode, trop éloignées de leur contrefactuel du matching évoqué précédemment sont identifiés puis ces réponses sont modifiées par hot-deck sur le contrefactuel afin d'annuler l'effet de mesure.

L'objectif de cette étude est alors de regarder l'impact de ces deux méthodes d'agrégation sur les estimations de paramètres de modélisations économétriques incluant des variables sujettes à des effets de mesure.

Il s'agira de discuter de la modélisation d'une variable présentant un effet de mesure à l'aide de variable sans effet de mesure, en fonction de la méthode d'agrégation. En particulier, l'introduction à la modélisation de la variable indicatrice précisant le mode de collecte sera analysée de façon approfondie : permet-elle à elle seule de saisir l'effet de mesure dans le cas d'une agrégation simple ? Permet-elle de voir un effet de mesure résiduel dans le modèle après correction par imputation ? L'estimation des paramètres du modèle est-elle sensible à la méthode d'agrégation ?

Dans un second temps, l'analyse se portera sur l'ajout, dans la modélisation, des variables ayant servi à la modélisation du score de propension qui a permis le matching, c'est-à-dire celles captant l'effet de sélection. Réduisent-elles l'effet de mode présent dans le modèle ?

Il sera ensuite discuté de l'effet de variables présentant des effets de mesure en tant que variables explicatives lors de la modélisation d'une variable non sujette à un effet de mesure, en tant que facteur isolé, ou en interaction avec la variable de mode de collecte.

La présentation se conclura par une discussion sur la mise à disposition des données aux chargés d'études : les informations à livrer, ainsi que les recommandations d'utilisation du jeu de données selon la méthode d'agrégation retenue et selon les objectifs d'utilisation des données (statistique descriptive ou modélisation)

Bibliographie

[1] Barret C., Dzikowski C., « La collecte par Internet est-elle l'avenir des enquêtes Génération du Céreq ? », *12èmes Journées de Méthodologie Statistique*, 2015.

[2] Barret C., Dzikowski C., « Evaluation d'un protocole multimode avec échantillon embarqué et agrégation des données en présence d'effet de mesure », *9e colloque francophone sur les sondages,* 2016

[3] Legleye, S., « Effets de sélection, imputations et effets de mode : les dernières tendances en matière de multimode », *Séminaire de Méthodologie Statistique,* 2017.