Econométrie et Données d'Enquête : les effets de l'imputation de la non-réponse partielle sur l'estimation des paramètres d'un modèle économétrique

C.Charreaux[1], **C. Favre-Martinoz** [2], H.Harle [1], R.Le Saout[3], P.-A. Robert[1]

[1] *Ecole Nationale de la Statistique et de l'Administration Economique*
[2] *Direction de la Méthodologie et de la Coordination Statistique et Internationale, INSEE*
[3] *Ecole Nationale de la Statistique et de l'Analyse de l'Information*

9^e Colloque Francophone sur les sondages, Gatineau
13 octobre 2016

Objectifs de la présentation

- Étudier les effets de l'imputation (simple) de la non-réponse partielle sur les analyses multivariées de variables quantitatives à l'aide de simulations, en liant théorie économétrique et théorie des sondages.
- Mise en garde sur la diffusion de données imputées à l'aide de méthodes utilisées pour des paramètres de population finie pour des études économétriques.
- En général, on a recours à des méthodes d'imputation simple, i.e pour ajuster la distribution marginale d'une variable d'intérêt $\rightarrow$ destruction des corrélations entre variables, car les distributions conjointes des variables d'intérêt sont modifiées
- Question : Le traitement de la non-réponse partielle introduit-il de l'endogénéité dans le modèle économétrique et crée t-il un biais supplémentaire ?

La non-réponse partielle

- Non-réponse partielle : une partie du questionnaire n'est pas renseignée. Plusieurs causes possibles :
  - l'individu a refusé de répondre à certaines questions de l'enquête (ces questions abordant par exemple des thèmes sensibles comme le revenu),
  - l'individu sélectionné ne comprend pas la question,
  - les réponses sont incohérentes et par conséquent invalidées,
  - l'enquêteur n'a pas compris la réponse de l'enquêté,
  - l'individu sélectionné abandonne au cours de l'enquête (enquête en plusieurs visites, carnet,....).

Méthodes d'imputation utilisées pour traiter la non-réponse partielle

- Les méthodes couramment utilisées pour traiter la non-réponse partielle sont les suivantes (entre parenthèses, nous indiquons les abréviations utilisées par la suite) :
  - imputation par la moyenne (Moyenne)
  - imputation par Hot-Deck (HotDeck)
  - imputation par Hot-Deck stratifié (HotDeckStrat)
  - imputation par plus proche voisin (KNN)
  - imputation par la régression (Regression)
  - Nous comparons dans la suite ces cinq méthodes classiques au cas où l'on supprime purement et simplement les unités non-répondantes appelé "Available Case".
- Ces méthodes permettent de corriger le biais pour l'estimation de paramètre univarié de population finie (total, moyenne).
- Ces techniques d'imputation marginale détruisent la corrélation pré-existante entre les variables.$\rightarrow$ Procédure d'imputation multivariée type Shao & Wang.

Modèle de superpopulation : modèle économétrique

- On considère quatre variables aléatoires continues : X, Z, Z' et ε et le modèle économétrique sans constante suivant :

$$Y = \beta \cdot X + \gamma \cdot Z' + \varepsilon,$$

  avec $\beta = 1$ et $\gamma = 0 \text{ ou } 1$.
- On suppose de plus que le terme d'erreur ε est indépendant des autres variables aléatoires afin de ne pas créer d'endogénéité.
- Les variables aléatoires X, Z et Z' sont supposées suivre des lois normales multivariées.
- Les variables aléatoires X et Z sont supposées corrélées entre elles mais non corrélées avec Z'.
- La non-réponse partielle est générée pour les variables X et Y à l'aide d'un mécanisme de non-réponse fonction des variables Z et/ou Z'.
- La variable Z' est inobservée.

Cas étudiés

- **Cas 1.** Exogénéité
  - **Cas 1.1** : ($\gamma = 0$ ou 1) La non-réponse est **fonction de** Z **uniquement**.
  - **Cas 1.2** : Le paramètre γ est supposé nul et la non-réponse fonction de Z et Z' $\rightarrow$ mauvaise spécification du mécanisme de non-réponse.

- **Cas 2.** Endogénéité
  - Le paramètre γ est supposé non nul et la non-réponse fonction **de** Z **et** Z' pour X uniquement (mais uniquement de Z pour Y, sinon la non-réponse est non ignorable).$\rightarrow$ Cette situation crée de la sélection et donc de l'endogénéité.
  - L'estimation du modèle économétrique sur les seules observations (X, Y) sans données manquantes sera biaisée. La question est alors de savoir si les méthodes d'imputation amplifient ou diminuent ce biais.

Protocole de simulation

- Une population U de 5000 individus avec deux variables d'intérêt X et Y, cette dernière étant générée à partir du modèle :

$$Y = \beta \cdot X + \gamma \cdot Z' + \varepsilon$$

avec Z' une variable auxiliaire inobservée et ε un terme d'erreur (étape 1).

- $\left(X, Z, Z'\right) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu} = (0, 2, 2)^\top$ et

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

- On effectue un sondage aléatoire simple sans remise (étape 2) de taille $n = 200$ ou $n = 500$ parmi $N = 5000$.
- La non-réponse partielle est générée selon un modèle logistique pour les variables X et Y (étape 3) de façon à obtenir un **taux de non-réponse moyen de 20%**.

Protocole de simulation

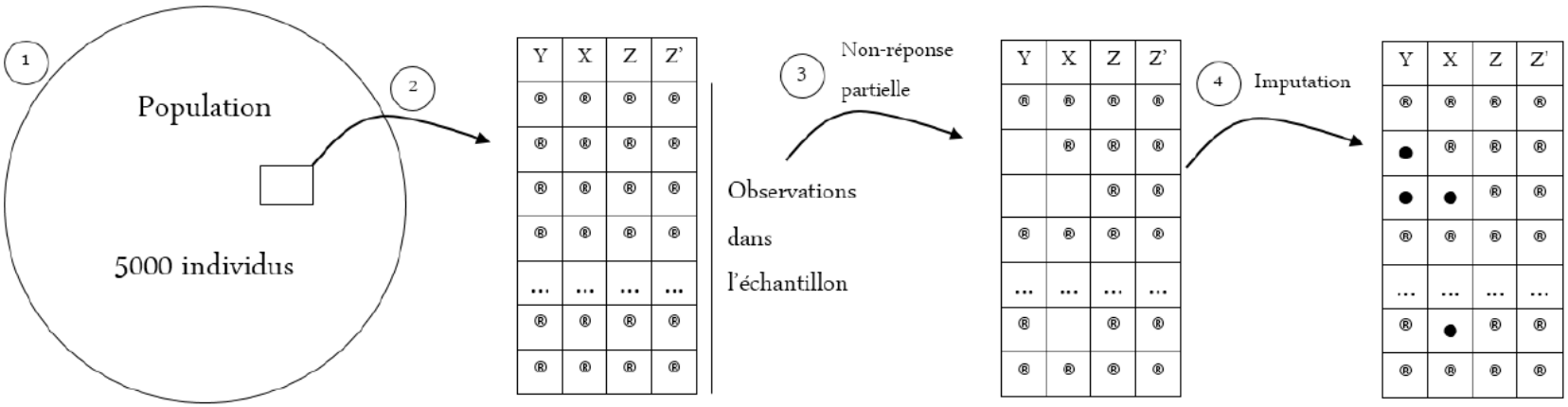

Figure: Protocole de simulation

Protocole de simulation

- Trois sources d'aléa sont ici présentes dans le processus générateur des données :
  - le modèle
  - le plan de sondage
  - la sélection des répondants
- Simplification : le plan de sondage étant ici un sondage aléatoire simple sans remise $\rightarrow$ le modèle économétrique qui tient au niveau de la population, tient également au niveau de l'échantillon. Cette étape n'engendre aucun biais dans l'estimation des paramètres du modèle.

Protocole de simulation

- La non-réponse est imputée par différentes méthodes, sous l'hypothèse d'un mécanisme de non-réponse fonction d'une variable auxiliaire parfaitement observée Z mais également de la variable auxiliaire inobservée Z'.
- L'objectif de nos simulations est d'observer l'effet de la méthode d'imputation sur l'estimation du paramètre de superpopulation β et de sa variance.

Les résultats évidents sur les paramètres de population finie

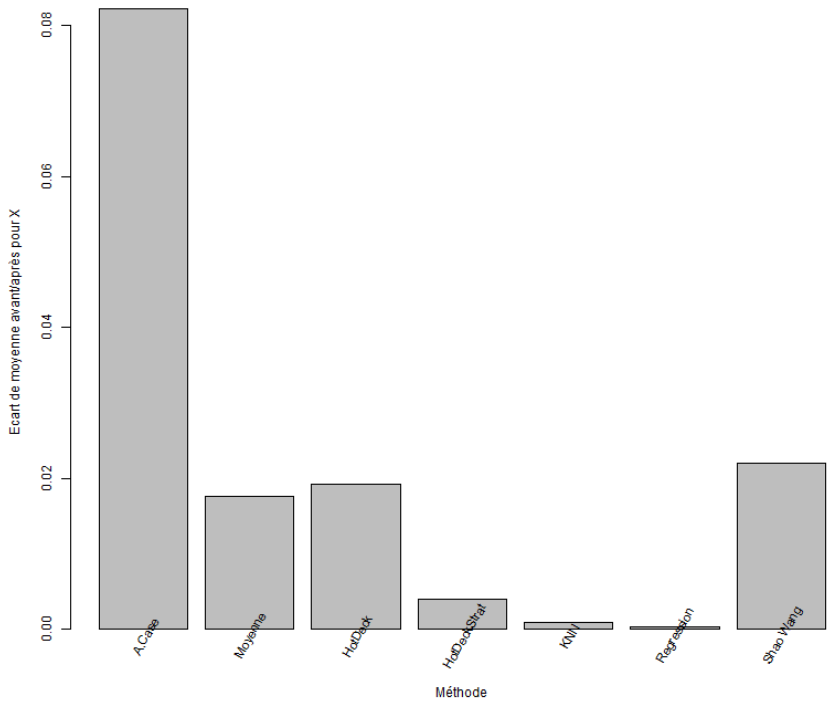

Figure: Effet de l'imputation sur la moyenne de X

Les résultats évidents sur les paramètres de population finie

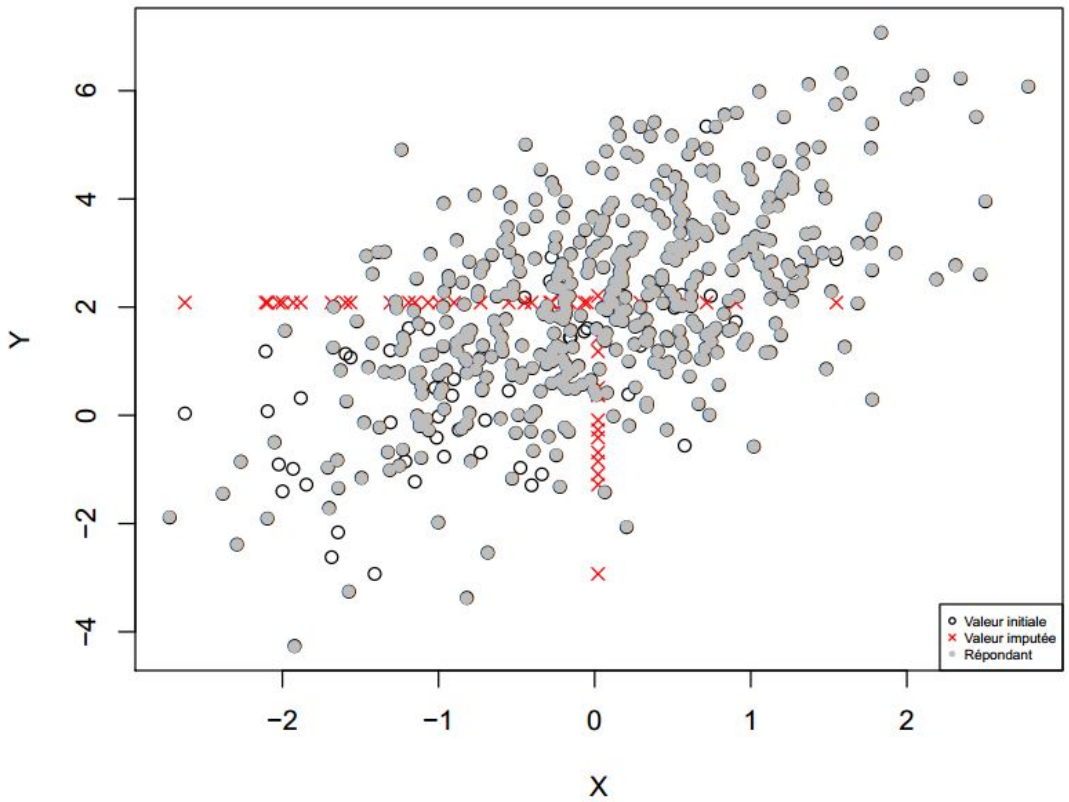

Figure: Effet de l'imputation par la moyenne sur les valeurs de Y et X

Les résultats évidents sur les paramètres de population finie

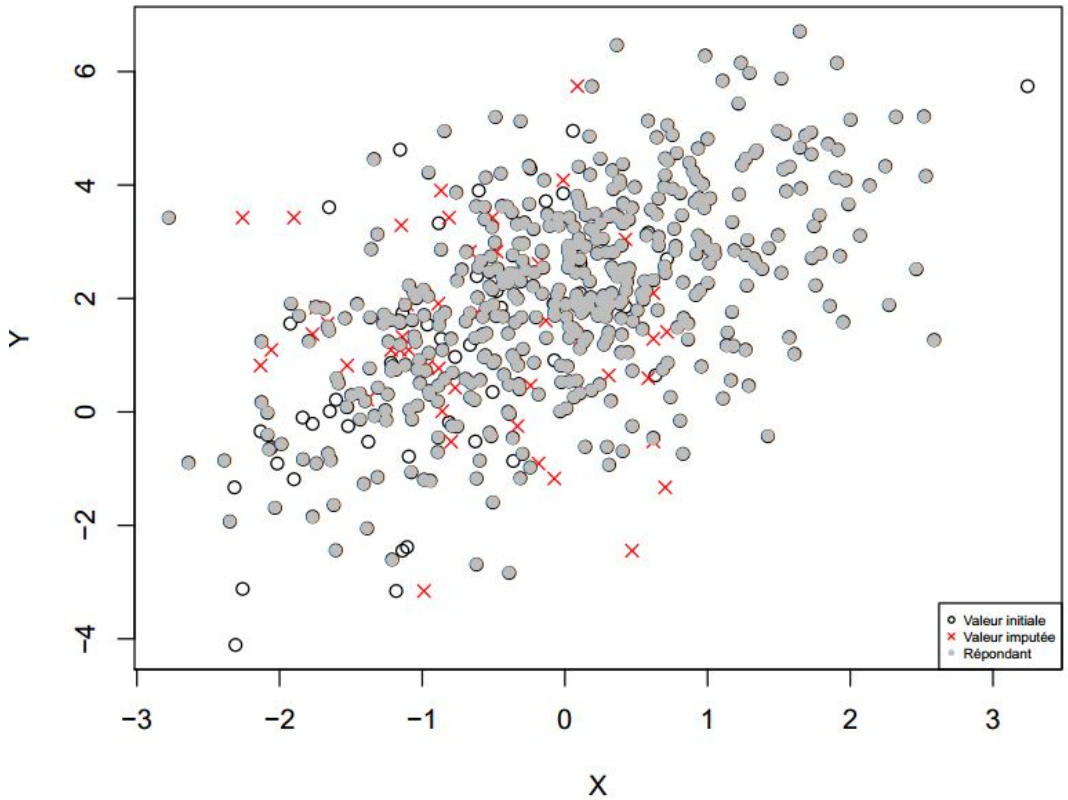

Figure: Effet de l'imputation par la méthode des plus proches voisins sur les valeurs de Y et X

Rappel des cas étudiés

Cas		γ	Var. de non-réponse (X)	Var. de non-réponse (Y)
Exogénéité	1.1	Non nul	Z	Z
	1.2	Nul	Z + Z'	Z +Z'
Endogénéité	2	Non nul	Z + Z'	Z

Figure: Tableau récapitulatif des cas étudiés

Résultats sur β

- A partir de K=5 000 simulations Monte-Carlo, nous avons estimé le biais relatif et la variance associée à l'estimateur des MCO calculé sur les répondants uniquement ou sur le jeu de données complété à l'aide d'une des six méthodes d'imputation :

$$BR_{MC}\left(\hat{\beta}^{MCO}\right) = \frac{E_{MC}\left(\hat{\beta}^{MCO}\right) - \beta}{\beta} \times 100,$$

où

$$E_{MC}\left(\hat{\beta}^{MCO}\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}^{MCO}(k),$$

et

$$Var_{MC}\left(\hat{\beta}^{MCO}\right) = E_{MC}\left\{\left[\hat{\beta}^{MCO} - E_{MC}\left(\hat{\beta}^{MCO}\right)\right]^2\right\}.$$

Résultats sur β

- Pour une taille d'échantillon $n = 200$:

Méthode	Biais (%)			100 x Var($\hat{\beta}$)		
Cas	1.1	1.2	2	1.1	1.2	2
A.Case	0.50	0.04	-0.97	1.43	0.74	1.29
Moyenne	-14.46	-13.50	-17.65	1.28	0.81	1.09
HotDeck	-27.58	-25.37	-21.60	1.49	0.94	1.38
HotDeckStrat	-20.38	-19.54	-15.22	1.51	0.92	1.37
KNN	-16.73	-17.30	-13.62	1.63	0.87	1.47
Regression	-9.47	-9.57	-11.68	1.36	0.76	1.17
Shao Wang	0.88	3.87	4.47	2.41	0.94	2.11

Table: Etude du biais estimé dans les trois cas, n=200, 5000 simulations

Résultats sur β

- Pour une taille d'échantillon $n = 500$:

Méthode	Biais (%)			100 x Var($\hat{\beta}$)		
Cas	1.1	1.2	2	1.1	1.2	2
A.Case	-0.02	0.00	-1.38	0.52	0.27	0.51
Moyenne	-14.98	-13.41	-17.75	0.46	0.27	0.46
HotDeck	-28.15	-25.49	-21.70	0.58	0.33	0.58
HotDeckStrat	-21.31	-19.50	-15.80	0.57	0.33	0.57
KNN	-17.49	-16.94	-14.02	0.60	0.32	0.59
Regression	-9.82	-9.40	-11.87	0.46	0.27	0.47
Shao Wang	0.08	3.57	3.85	0.88	0.34	0.83

Table: Etude du biais estimé dans les trois cas, n=500, 5000 simulations

Interprétation des résultats

- Le premier constat est l'importance du biais estimé, qui peut atteindre 25% en moyenne (ex : HotDeck).
- Dans le cas présenté ici, la meilleure méthode pour limiter le biais sur β reste l'estimation par "Available Case".
- En utilisant non plus 200 mais 500 individus, soit un passage de 2% à 10% de la taille de la population, on constate une division par plus de 2 de la variabilité de $\hat{\beta}$.
- Le biais moyen estimé quant à lui conserve le même ordre de grandeur malgré une taille d'échantillon plus grande.
- Ce biais diminue lorsque le taux de non-réponse diminue.

Quelles sont les bonnes pratiques à adopter ?

- Le choix de la méthode d'imputation et son signalement n'est donc pas anodin pour l'analyse économétrique de données présentant de la non-réponse partielle.
- Si les drapeaux d'imputation sont disponibles, il est possible de revenir aux données bruts.
  - La question peut se poser alors de ne pas tenir compte des imputations effectuées.
  - Traiter la non-réponse partielle à l'aide de méthodes de sélection, d'identification partielle, d'un algorithme EM ou d'imputations multiples.
  - Ce choix reste bien sûr coûteux par rapport à l'utilisation directe de la base de données diffusée.

Merci de votre attention !