TRAITEMENT DE LA NON-RÉPONSE NON-IGNORABLE PAR CALAGE GÉNÉRALISÉ

Une simulation à partir de l'enquête Budget des Ménages au Luxembourg

Guillaume Osier

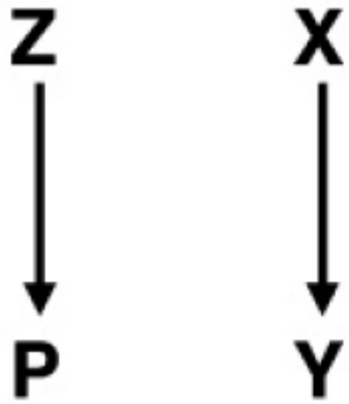
Institut National de la Statistique et des Etudes Economiques du Grand-duché de Luxembourg (STATEC)

Division des Statistiques Sociales

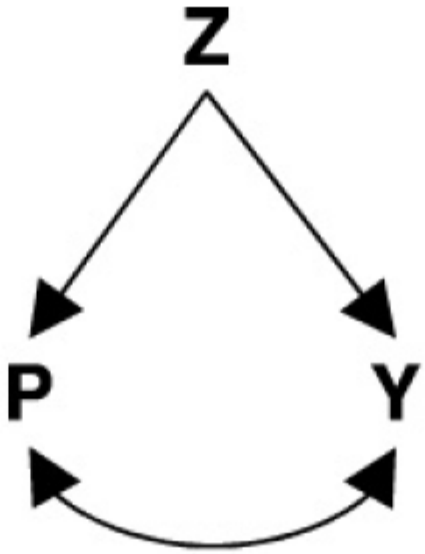
Objectif de la présentation

- Tester la méthode du calage généralisé, telle qu'elle est implantée dans la macro Calmar 2, comme moyen de traitement de la non-réponse lorsque celle-ci dépend des variables d'intérêt de l'enquête (mécanisme non-ignorable)
- On a réalisé pour cela une simulation à partir des données de l'enquête Budget des Ménages (EBM) au Luxembourg
- Cette enquête, comme toutes les enquêtes budget en général, souffre d'importants problèmes de non-réponse

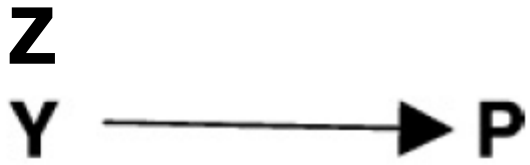
Les mécanismes de non-réponse



Causes « séparées »



Causes « partagées »



Relation causale avec les variables d'intérêt (mécanisme non-ignorable)

Mécanismes non-ignorables

- Certainement très répandus (enquêtes sur les revenus, le patrimoine, la victimation…)
- Comme, par construction, les variables d'intérêt de l'enquête ne sont connues que sur les unités répondantes, on ne pourra pas construire un modèle permettant d'estimer les probabilités de réponse
- Les méthodes de calage constituent une alternative pour traiter ces cas de non-réponse

Calage simple (1/2)

On modifie « à la marge » les poids d'échantillonnage $d = \{d_k, k \in r\}$ de façon à ce que les totaux de calage soient estimées de façon exacte à partir des nouvelles pondérations :

$$\omega = \{\omega_k, k \in r\} = Argmin_{\rho=\{\rho_k, k \in r\}} \sum_{k \in r} D(\rho_k, d_k)$$

Sous la contrainte de calage :

$$\sum_{k \in r} \omega_k x_k = X$$

Calage simple (2/2)

Les poids « calés » s'écrivent : $\omega_k = d_k F(x'_k \lambda)$, avec

$$\sum_{k \in r} d_k F(x'_k \lambda) x_k = X$$

On montre que si les variables de calage sont corrélées avec la probabilité de réponse, alors le calage simple permet de réduire le biais dû à la non-réponse

Problème: il faut connaître les totaux de calage

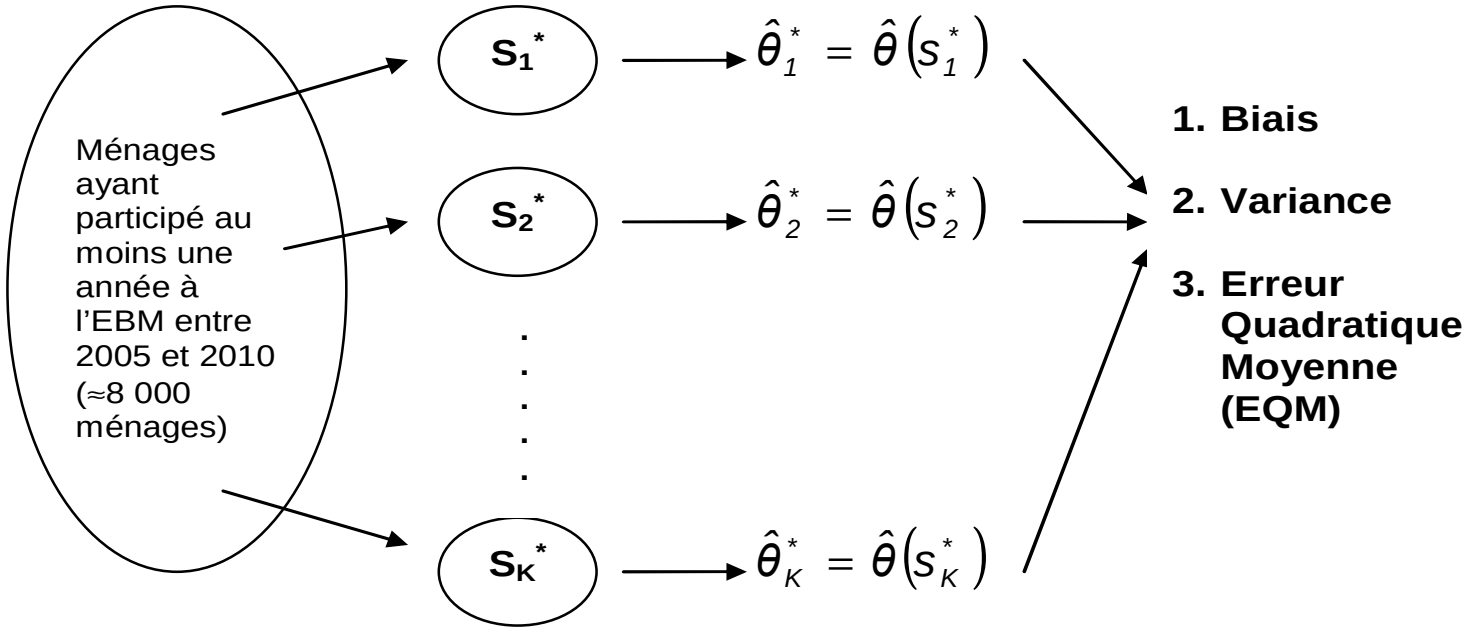
Calage généralisé

Les poids « calés » s'écrivent : $\omega_k = d_k F(z'_k \lambda)$, avec

$$\sum_{k \in r} d_k F(z'_k \lambda) x_k = X$$

Les variables z permettent de corriger la non-réponse, tandis que les variables de calage x réduisent la variance d'échantillonnage. La seule condition sur les variables de non-réponse est de connaître leurs valeurs sur les unités répondantes (mais plus besoin du total)

Les différentes étapes de la simulation



Tirage de K = 2000 réplications selon un mécanisme non-ignorable (tirage « comme dans l'EBM »)

Estimation du paramètre-cible θ à partir des données de chaque réplication

Construction du mécanisme non-ignorable

- Chacune des 2000 réplications est sélectionnée selon un tirage de Poisson
- La probabilité de tirage d'un ménage correspond à sa probabilité de réponse à l'EBM. Pour l'estimer, on va postuler un modèle logistique
- Variables du modèle : âge, genre et nationalité de la personne de référence, taille du ménage (nb de personnes) et logarithme de la dépense du ménage (variable d'intérêt)

Les différentes pondérations (1/2)

- Taux de réponse uniforme : $\omega_k = \frac{1}{\bar{\theta}}$
- Calage simple en une étape : $\omega_k = F(x_k\lambda)$, avec $\sum_{k\in r} F(x_k\lambda)x_k = X$
- Calage généralisé : $\omega_k = F(z_k\lambda)$, avec $\sum_{k\in r} F(z_k\lambda)x_k = X$

Le calage généralisé permet d'utiliser les variables d'intérêt de l'enquête pour corriger la non-réponse. Ces variables ne sont connues que sur les unités répondantes

Les différentes pondérations (2/2)

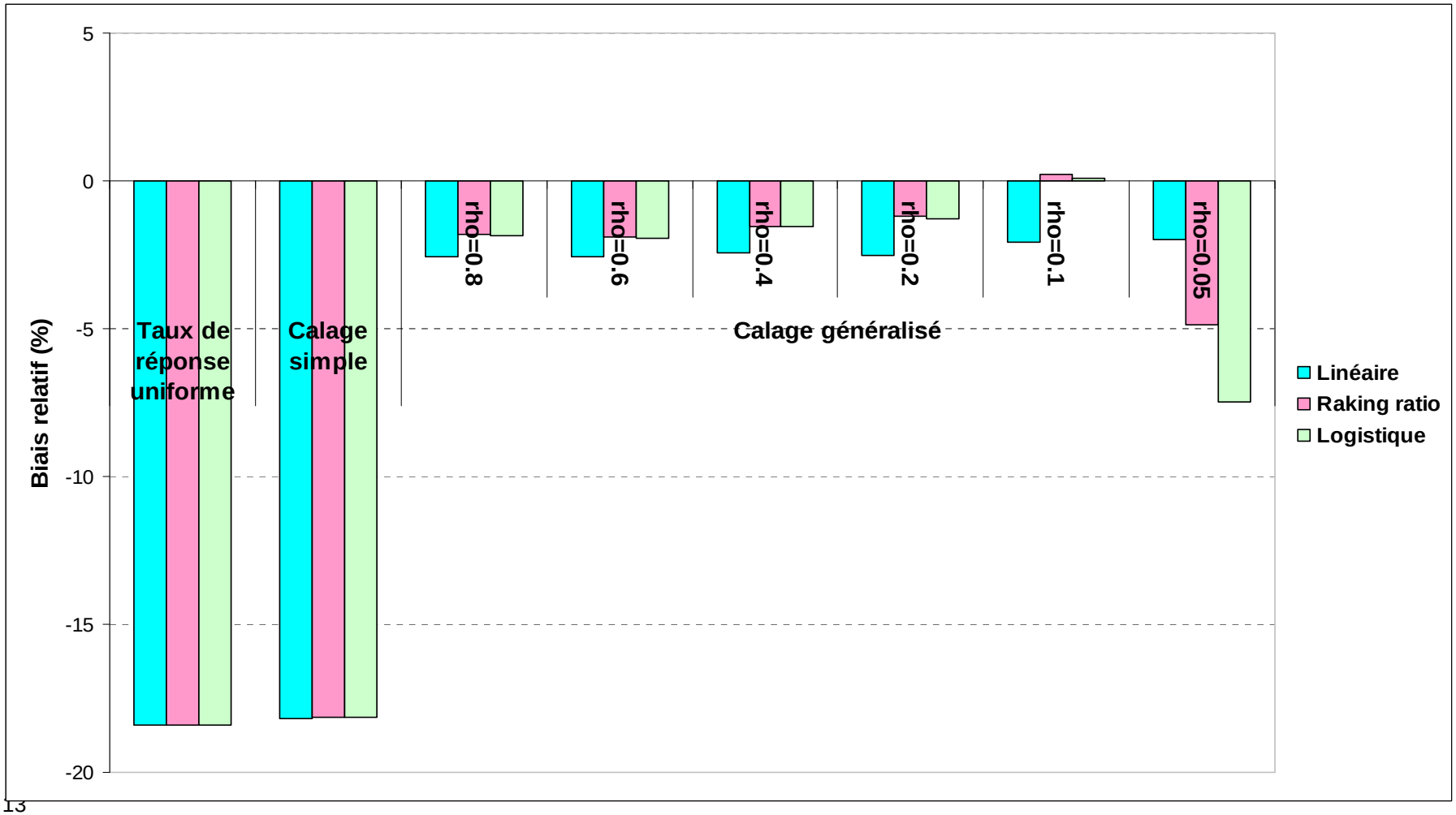
- **Variables utilisées pour le calage simple:** âge, genre et nationalité de la personne de référence, taille du ménage
- **Variables utilisées pour le calage généralisé:**
  - *Variables de non-réponse:* âge, genre et nationalité de la personne de référence, taille du ménage **+ dépense de consommation du ménage**
  - *Variables de calage:* âge, genre et nationalité de la personne de référence, taille du ménage **+ une variable de calage fictive ayant une corrélation ρ avec la dépense *t* du ménage**

La macro Calmar 2

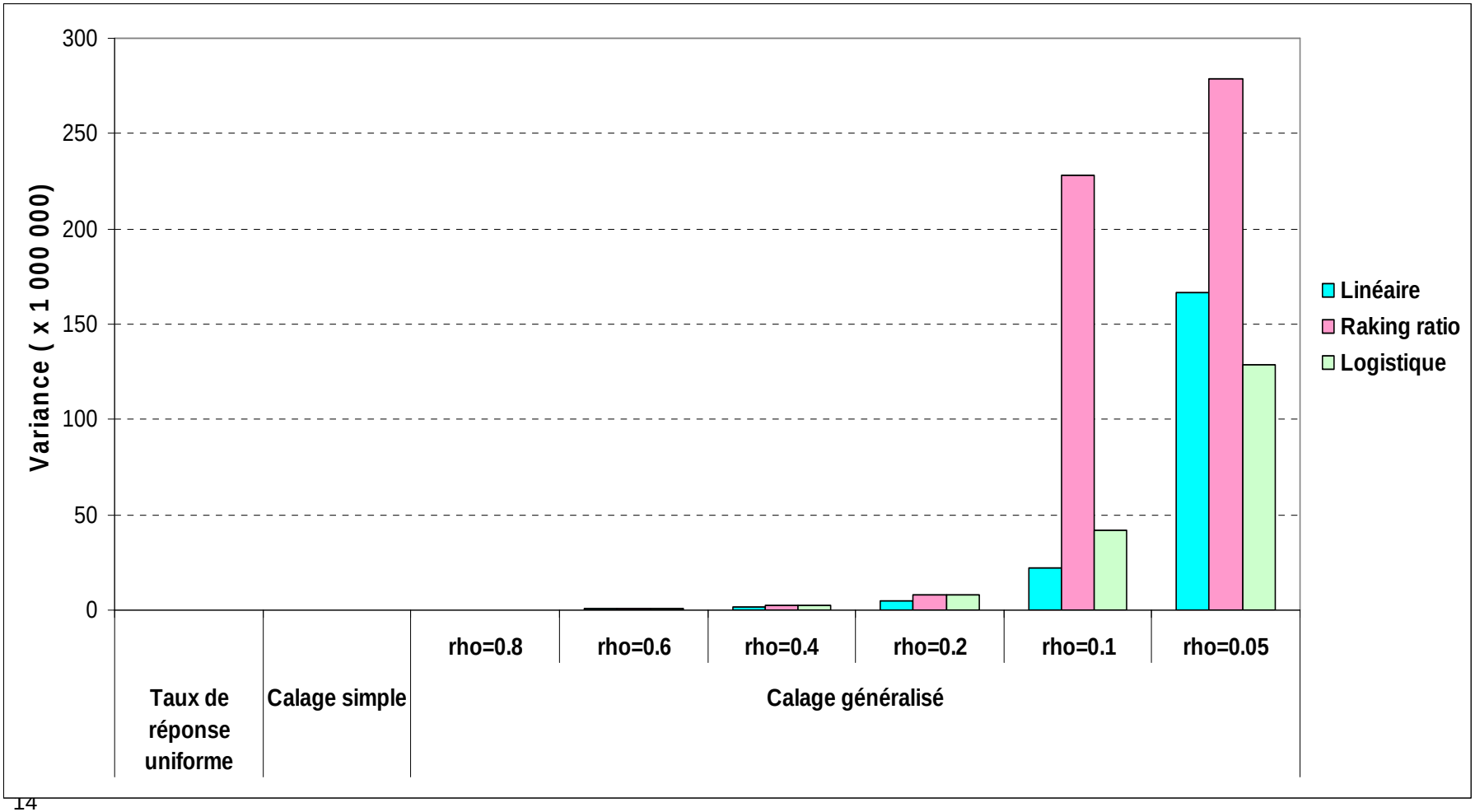
- Nouvelle version de la macro Calmar
- Prend en charge la technique du calage généralisé
- Outil utilisé dans la simulation pour calculer les pondérations
- Mise en œuvre très simple:
  - Table des marges
  - Paramètre NONREP = OUI

VAR	N	R	MAR1	
age1	0	0	568	Var. de calage
age2	0	0	3872	
age3	0	0	2153	
D60	0	0	7948	
age1	0	1	.	Var. de NR
age2	0	1	.	
age3	0	1	.	
t_00	0	1	.	

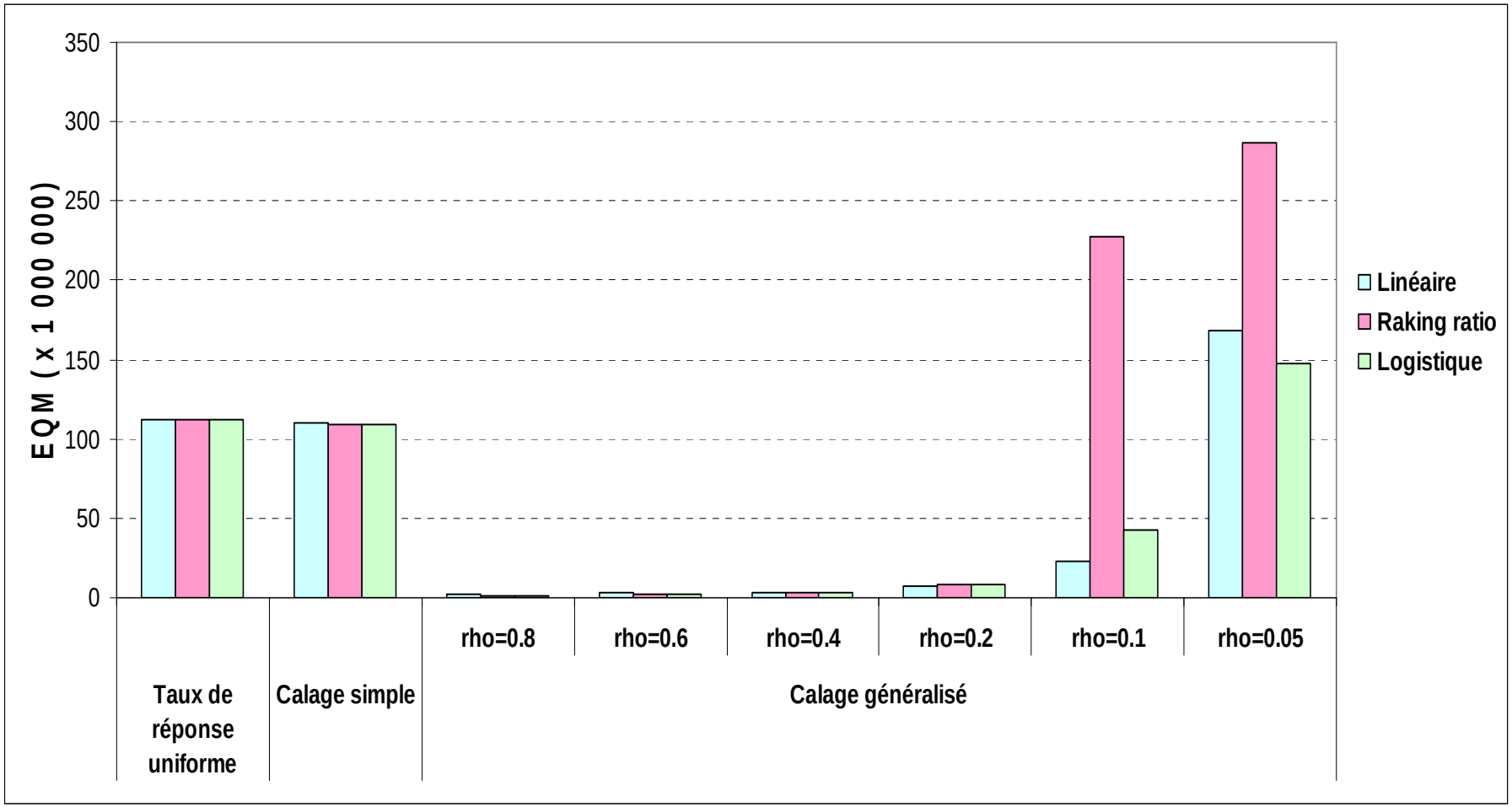
Résultat (cas non-ignorable): Biais



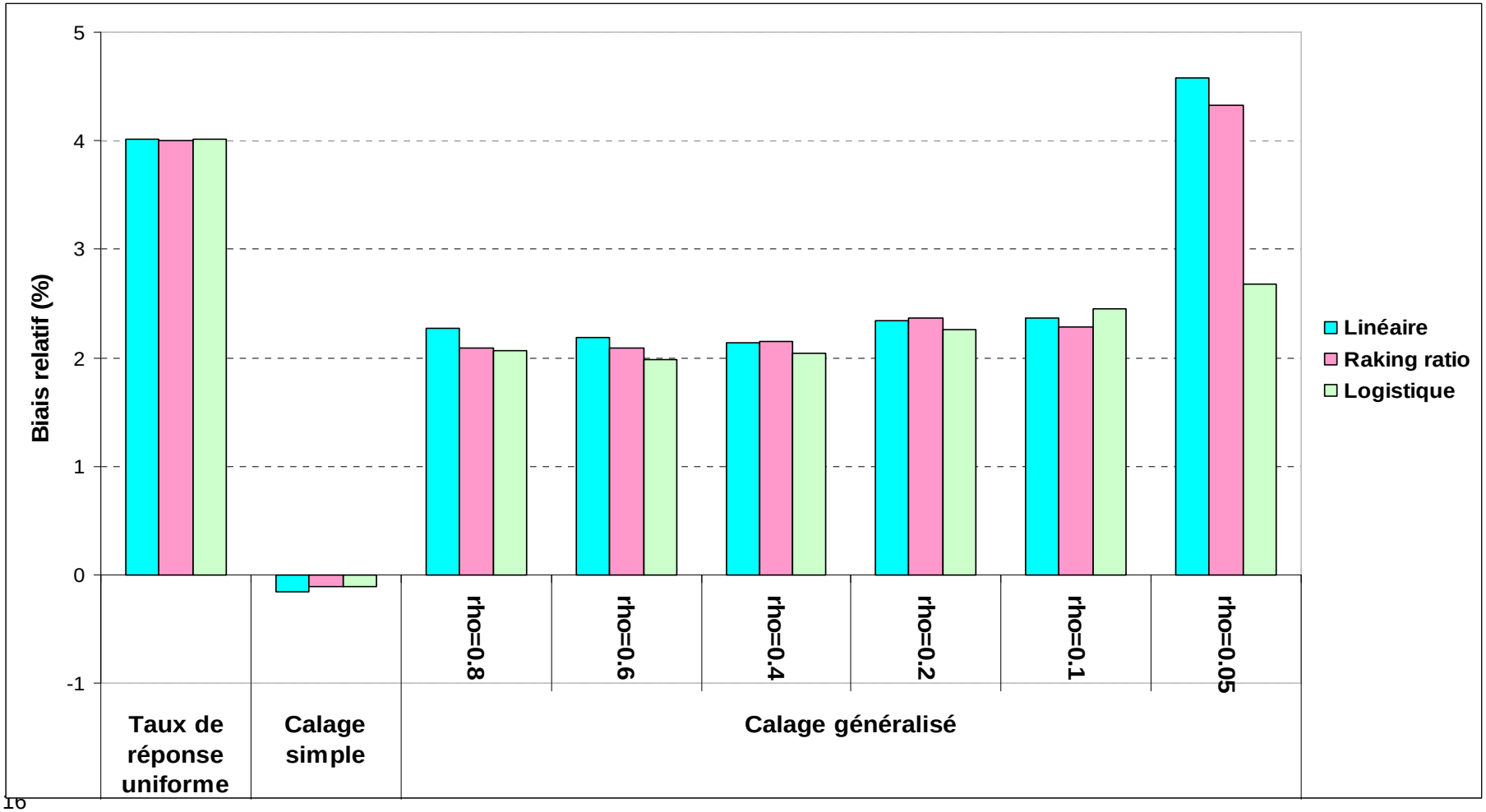
Résultat (cas non-ignorable): Variance



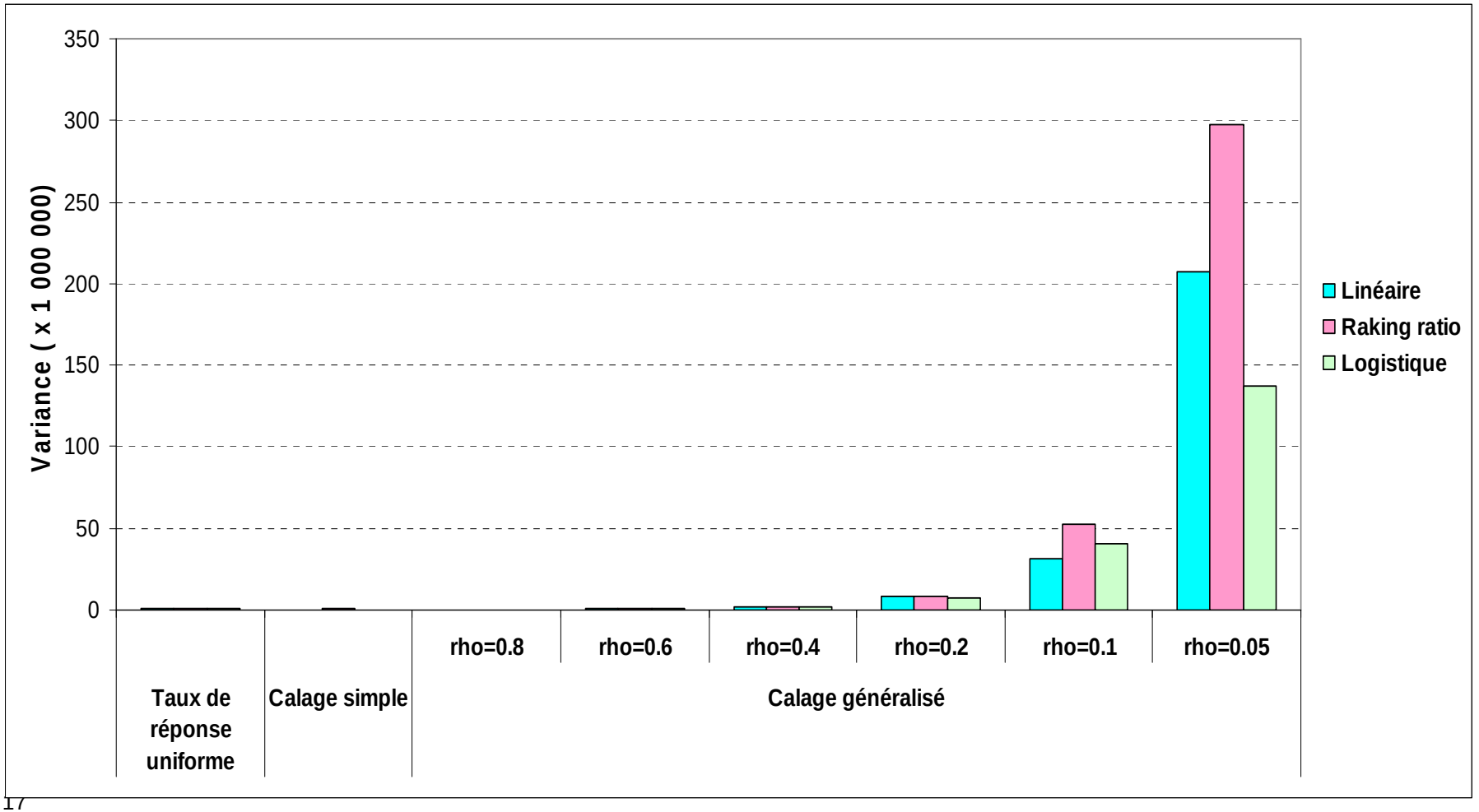
Résultat (cas non-ignorable): EQM



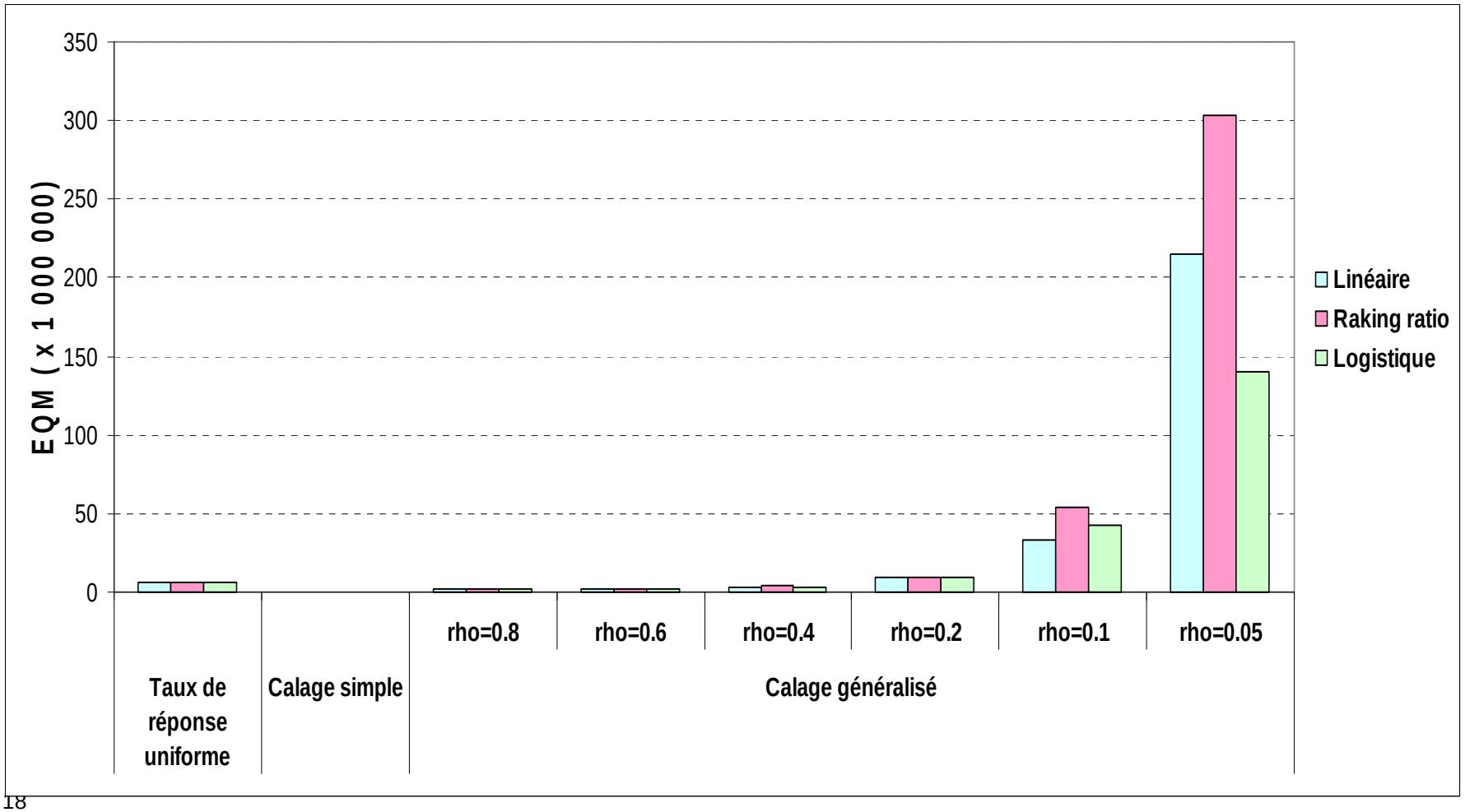
Résultat (cas ignorable): Biais



Résultat (cas ignorable): Variance



Résultat (cas ignorable): EQM



Conclusion (1/3)

Les résultats de cette simulation montrent que le calage généralisé peut réduire sensiblement le biais dû à la non réponse, en particulier lorsque celle-ci dépend des variables d'intérêt de l'enquête (mécanisme non-ignorable). Il s'agit là d'un résultat assez remarquable, qu'on ne peut pas atteindre avec un calage simple ou en estimant explicitement les probabilités de réponse. Le prix à payer pour ce résultat est une dégradation de la variance d'échantillonnage, d'autant plus importante que les variables de non-réponse sont peu corrélées avec les variables de calage. L'estimateur calé devient donc de plus en plus instable.

Conclusion (2/3)

Le calage généralisé est donc une méthode à double tranchant, qu'il faut employer avec un minimum de discernement. Il faut commencer par se convaincre que la non-réponse est influencée par un certain nombre de caractéristiques qui sont observées uniquement sur les unités répondantes. Cette étape est délicate. Elle constitue un pari (certains parlent même d'un « acte de foi ») dans la mesure où les techniques statistiques ne permettent pas de justifier une telle assertion (il faudrait faire des tests cognitifs en laboratoire). Sans autres arguments, il faut donc croire à l'hypothèse que l'on fait.

Conclusion (3/3)

Même si l'on est convaincu du lien entre la non-réponse et un certain nombre de variables d'intérêt de l'enquête, il faut aussi se demander quelle est l'importance du biais que cela génère. Si le biais est relativement faible, voire inexistant, on ne gagnera finalement pas grand-chose à le corriger par calage généralisé. On risque même de dégrader la qualité d'ensemble de l'estimateur puisque le gain (faible) lié à la réduction du biais sera perdu en raison de la dégradation de la variance. On voit donc qu'il y a un arbitrage à faire non seulement sur l'existence d'un biais mais aussi sur son importance, et que là aussi cet arbitrage relève du pari, de l'acte de foi.