CSTB
le futur en construction

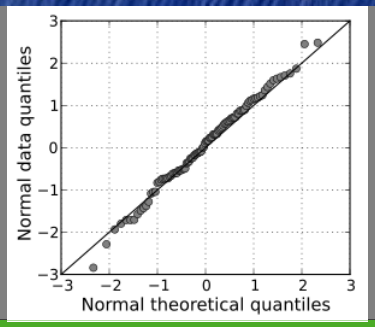
UNIVERSITÉ DE NANTES

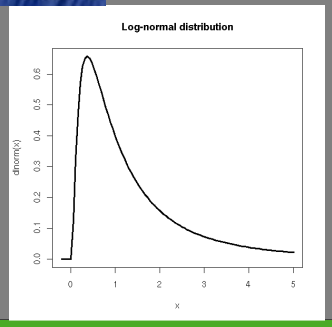
EA 4275
BIOSTATISTIQUE,
PHARMACOÉPIDÉMIOLOGIE,
ET MESURES SUBJECTIVES EN SANTÉ

$$y_{ij} = \sum_{p=0}^{8} \beta_{pj} x_{pij} + \sum_{v=9}^{13} \beta_{vj} z_{v-8ij} + \varepsilon_{ij}, \quad (14)$$

$$\beta_{pj} = \gamma_{0p} + u_{pj} \text{ for } p = 0, 1, \ldots, 8, \quad (15)$$

$$\beta_{vj} = \gamma_{0v} + u_{vj} \text{ for } v = 9, 10, \ldots, 13, \quad (16)$$





7ème colloque francophone sur les sondages, Ensai, Bruz, 7 novembre 2012 >Jean-Paul LUCAS

Modélisation multi-niveaux sur données d'enquête

Impact des poids de niveau 2 introduits dans la pseudo- vraisemblance

Introduction

- Modélisation multi-niveaux (MLM) sur données d'enquête : domaine de recherche récent (fin des années 1990), non encore abouti (noms du domaine : Pfeffermann, Asparouhov, Rabe-Hesketh et Skrondal, …)

- Jargon : Degrés $\longrightarrow$ Niveaux et incrémentation inversée (degré 1 devient niveau 2, degré 2 devient niveau 1….)

- Recherches surtout sur modèle à 2 degrés, sur données simulées et focalisation sur les poids des unités de niveau 1.

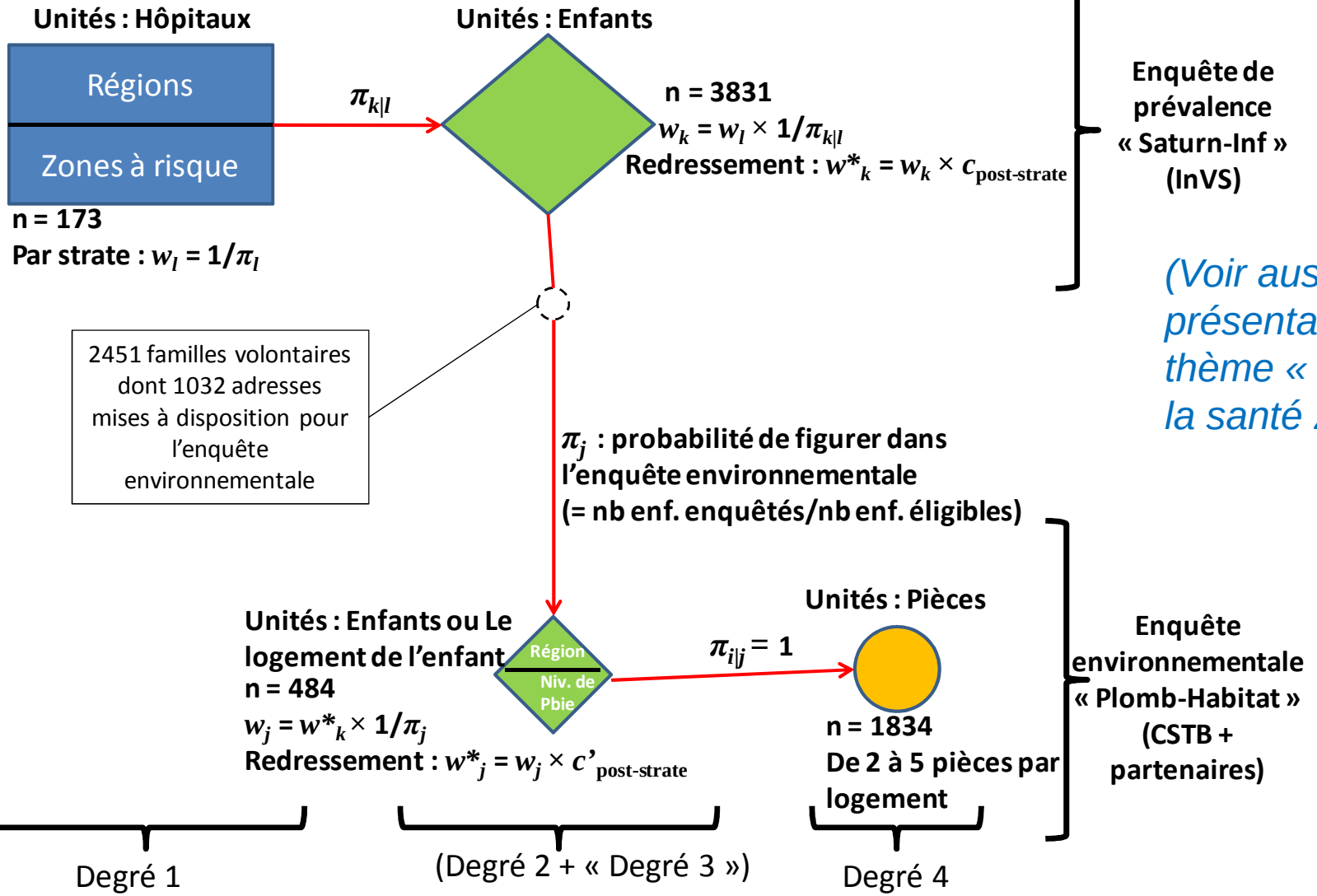
- Il faut connaitre les probabilités de sélection (conditionnelles) des unités à chaque niveau

Notre problématique

- On souhaite construire un modèle à but d'estimer les effets de sources en plomb (Pb) dans la contamination des poussières intérieures au sol dans les logements français.
- On a des concentrations en Pb dans plusieurs pièces (niveau 1) dans un logement (niveau 2).
- Echantillon de 484 logements

➔ **Enquête Plomb-Habitat**

Plan de sondage



(Voir aussi la présentation faite hier en thème « 20. enquête sur la santé 2»).

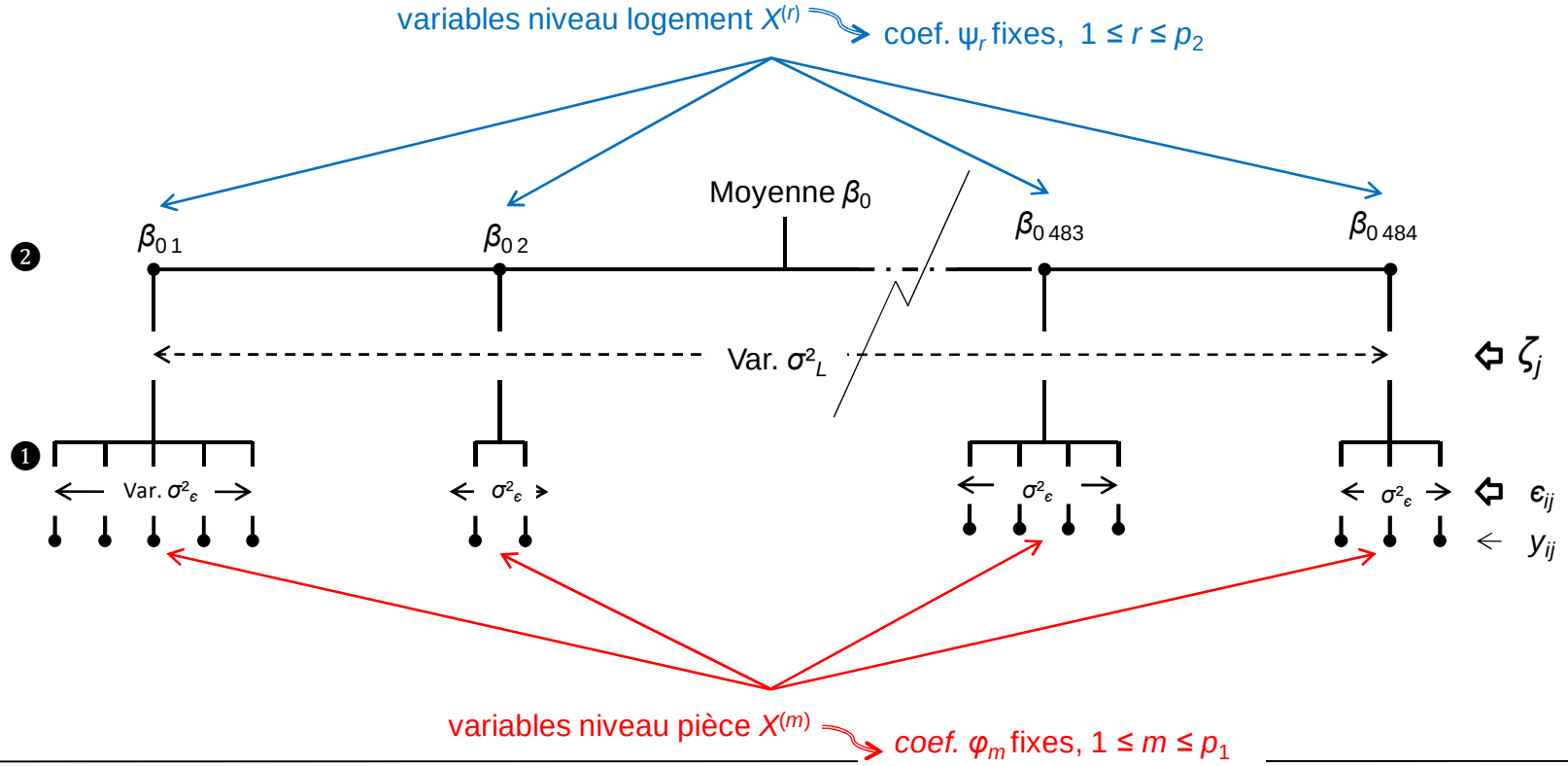
Modèle à 2 niveaux à intercepte aléatoire

Les niveaux hiérarchiques du modèle se décomposent :

Au niveau de la pièce *i* du logement *j* par : $y_{ij} = \beta_{0j} + \sum_{m=1}^{m=p_1} \varphi_m x_{ij}^{(m)} + \epsilon_{ij}$

Au niveau du logement *j* par : $\beta_{0j} = \beta_0 + \zeta_j + \sum_{r=1}^{r=p_2} \psi_r x_j^{(r)}$

$$\epsilon_{ij} \overset{iid}{\rightarrow} N(0, \sigma_c^2)$$
$$\zeta_j \overset{iid}{\rightarrow} N(0, \sigma_L^2)$$



Pseudo-vraisemblance

- Usual marginal log likelihood (without weights)

$$\log \prod_{j=1}^{n^{(2)}} \int \underbrace{\left\{ \prod_{i=1}^{n_j^{(1)}} f(y_{ij}|\zeta_j) \right\}}_{\Pr(\mathbf{y}_j|\zeta_j)} g(\zeta_j)\, \mathrm{d}\zeta_j = \sum_{j=1}^{n^{(2)}} \log \int \exp \left\{ \sum_{i=1}^{n_j^{(1)}} \log f(y_{ij}|\zeta_j) \right\} g(\zeta_j)\, \mathrm{d}\zeta_j$$

- Log pseudolikelihood (with weights)

$$\sum_{j=1}^{n^{(2)}} w_j \log \int \exp \left\{ \sum_{i=1}^{n_j^{(1)}} w_{i|j} \log f(y_{ij}|\zeta_j) \right\} g(\zeta_j)\, \mathrm{d}\zeta_j$$

- Note: need $w_j = 1/\pi_j$, $w_{i|j} = 1/\pi_{i|j}$; cannot use $w_{ij} = w_{i|j} w_j$

(Diapo issue de « Rabe-Hesketh, *Multilevel Modeling of Complex Survey Data*, 2007 West Coast Stata Users Group Meeting »)

Pseudo-vraisemblance - suite

- Log pseudolikelihood (with weights)

$$\sum_{j=1}^{n^{(2)}} w_j \log \int \exp \left\{ \sum_{i=1}^{n_j^{(1)}} w_{i|j} \log f(y_{ij}|\zeta_j) \right\} g(\zeta_j)\, \mathrm{d}\zeta_j$$

Lorsque les niveaux du modèle ne « match » pas les degrés du plan : quels poids w_j doivent être utilisés ?

Exemple : Données PISA (Programme international pour le suivi des acquis des élèves) dans Rabe-Hesketh S. and Skrondal A, 2006. Multilevel modelling of complex survey data. J.R. Statist. Soc. A, 169, pp 805-827.

Que faire sur nos données ?

On connait la construction de tous les poids : basé sur les probabilités conditionnelles, ou bien poids finaux sans redressement, poids finaux avec redressement etc. Avec poids au niveau 1 égal à 1.

9 scénarios listés, 6 en modèle à 2 niveaux, 3 en modèle à 3 niveaux.

Scenario	**w_1**	**w_2**	**w_3**	**w_4**	**w_5**	**w_6**	**w_1+** psu	**w_5 +** psu	**w_6** + psu
Nb de niveaux	2	2	2	2	2	2	3	3	3
Poids au plus haut niveau	-	-	-	-	-	-	$1/\pi_l$	1	$1/\pi_l$
Poids niveau 2	$1/\pi_j$	$w_j = \tilde{w}_{kl} \times 1/\pi_j$ $\tilde{w}_j$		$1/(\pi_{kl} \times \pi_j)$	1	$1/(\pi_{k\|l} \times \pi_j)$	$1/\pi_j$	1	$1/(\pi_{k\|l} \times \pi_j)$
Poids niveau 1	1	1	1	1	1	1	1	1	1

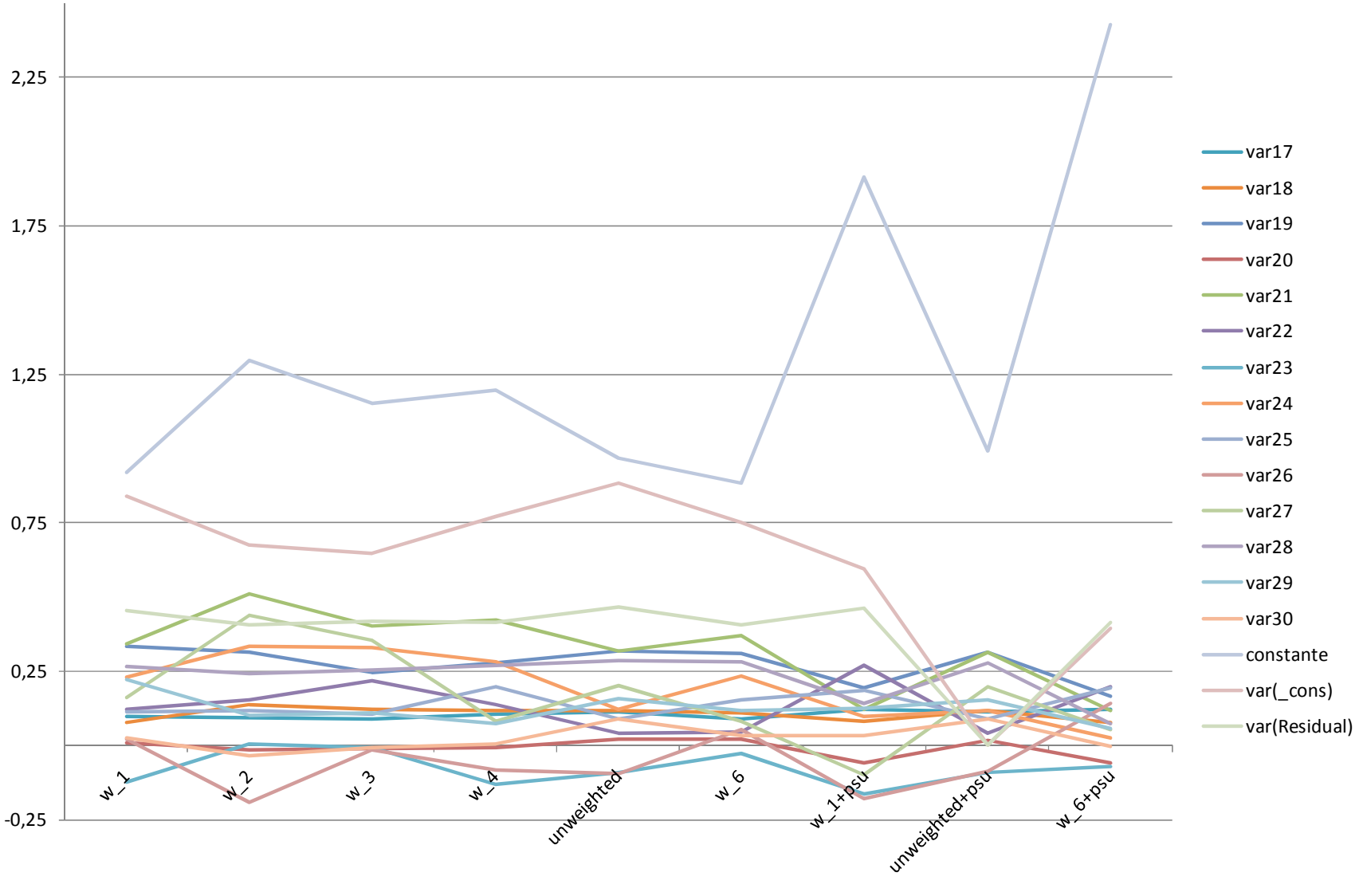
Logiciel utilisé

Fonction `xtmixed` qui gère depuis la version 12 les données d'enquête et les mises à l'échelle des poids des niveaux 1 recommandées par la littérature (uniquement en modélisation à 2 niveaux).

```
xtmixed Y list_of_covariates [pw = Pi_i_sachant_j]
|| ID_deg1_unit: , variance ml pweight(w_1)
```

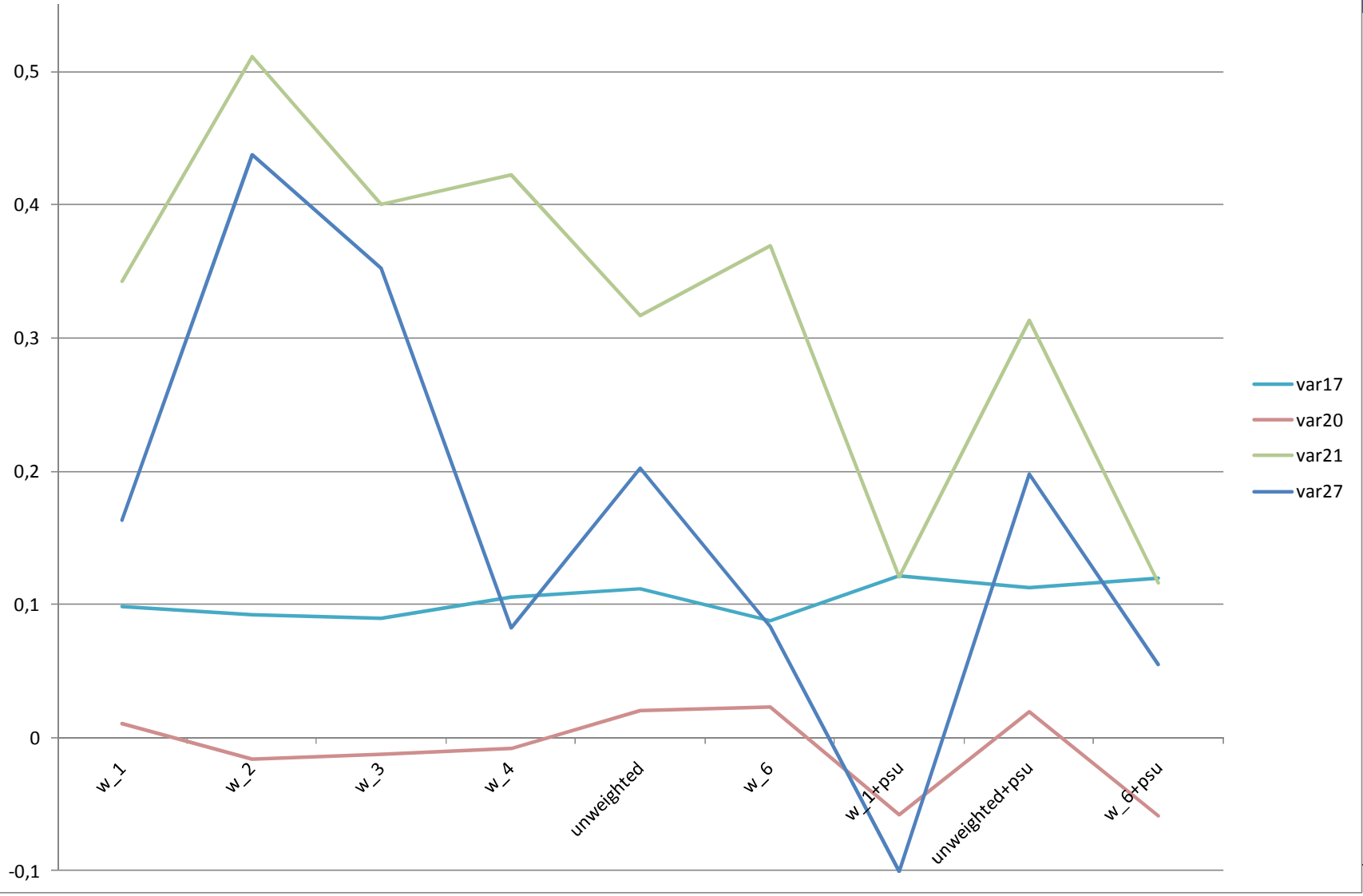
Impact sur les estimations

Estimation des paramètres du modèle selon le scénario
(variables sources en Pb uniquement + paramètres de covariance)

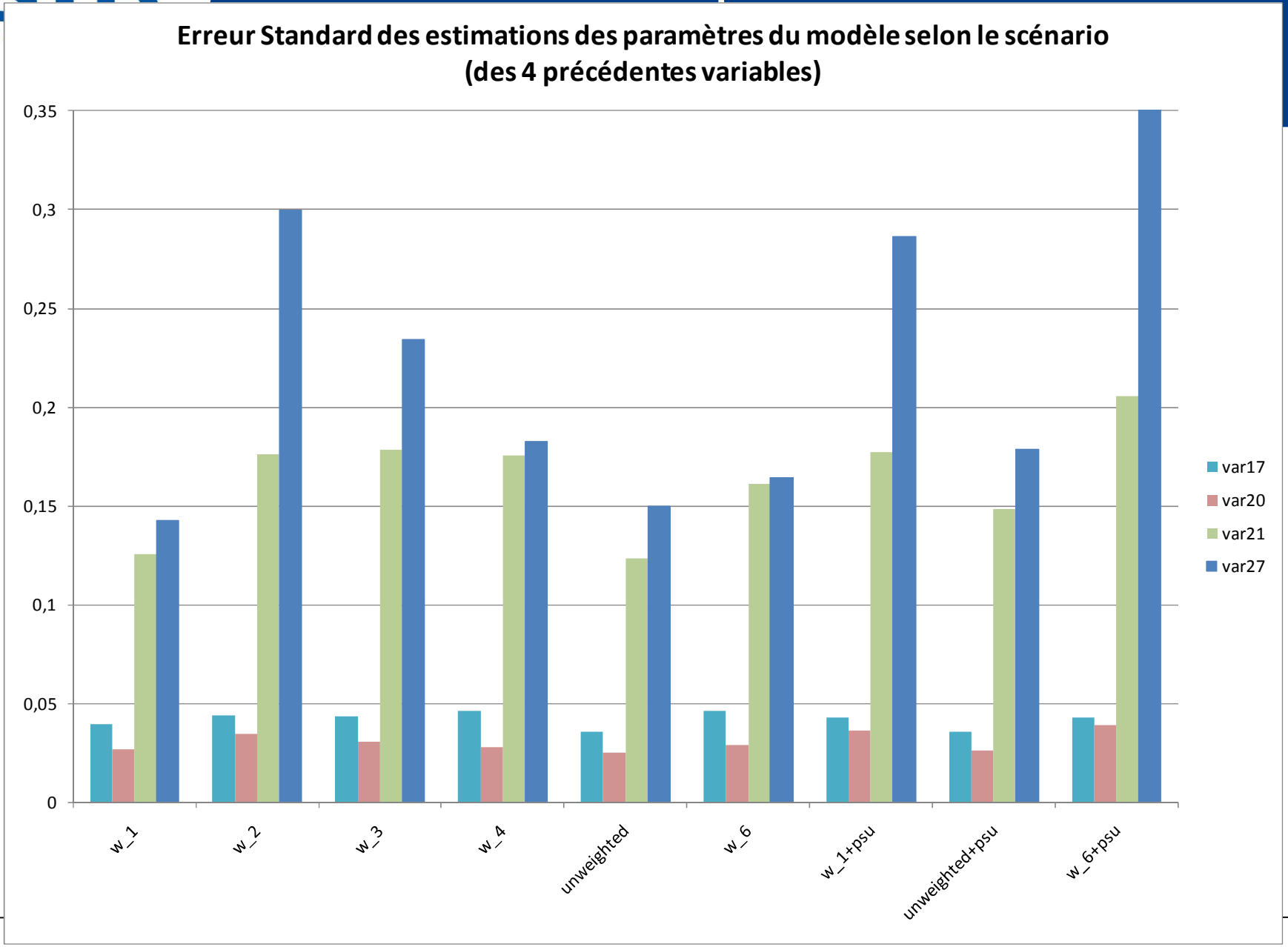


Impact sur les estimations

Estimation des paramètres du modèle selon le scénario
(4 variables résumant les comportements)



Impact sur les estimations



Comment choisir entre les scénarios ?

Stratégie pour répondre.

1) Générer une population de logements/pièces de taille *N*
2) Tirer 500 échantillons selon un plan de sondage analogue au notre
3) Estimer les paramètres selon les 9 scénarios
4) Comparer les paramètres estimés à leur vraies valeurs.

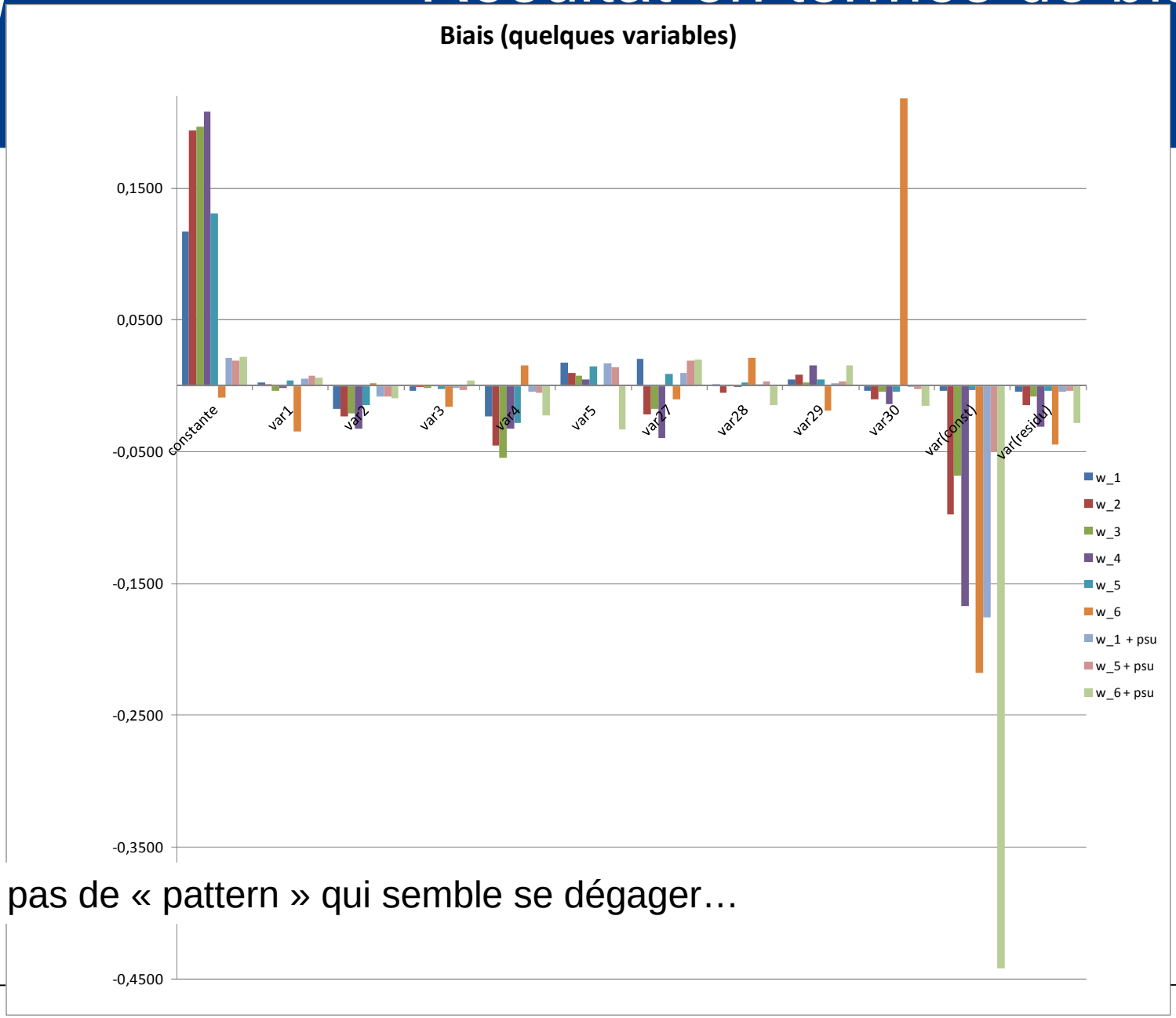
Génération d'une population

Génération et pas simulation car on utilise :

- les distributions des covariables du modèles estimées sur les données de notre enquête (en « *design-based* »)
- la base du recensement Insee 2006 : utilisation de variables auxiliaires le cas échéant, communes aux 2 fichiers
- *Y* générée à partir des valeurs des paramètres fixées selon les estimations obtenues pour le moment sur nos données selon les scénarios.

➔ Données non totalement « artificielles »

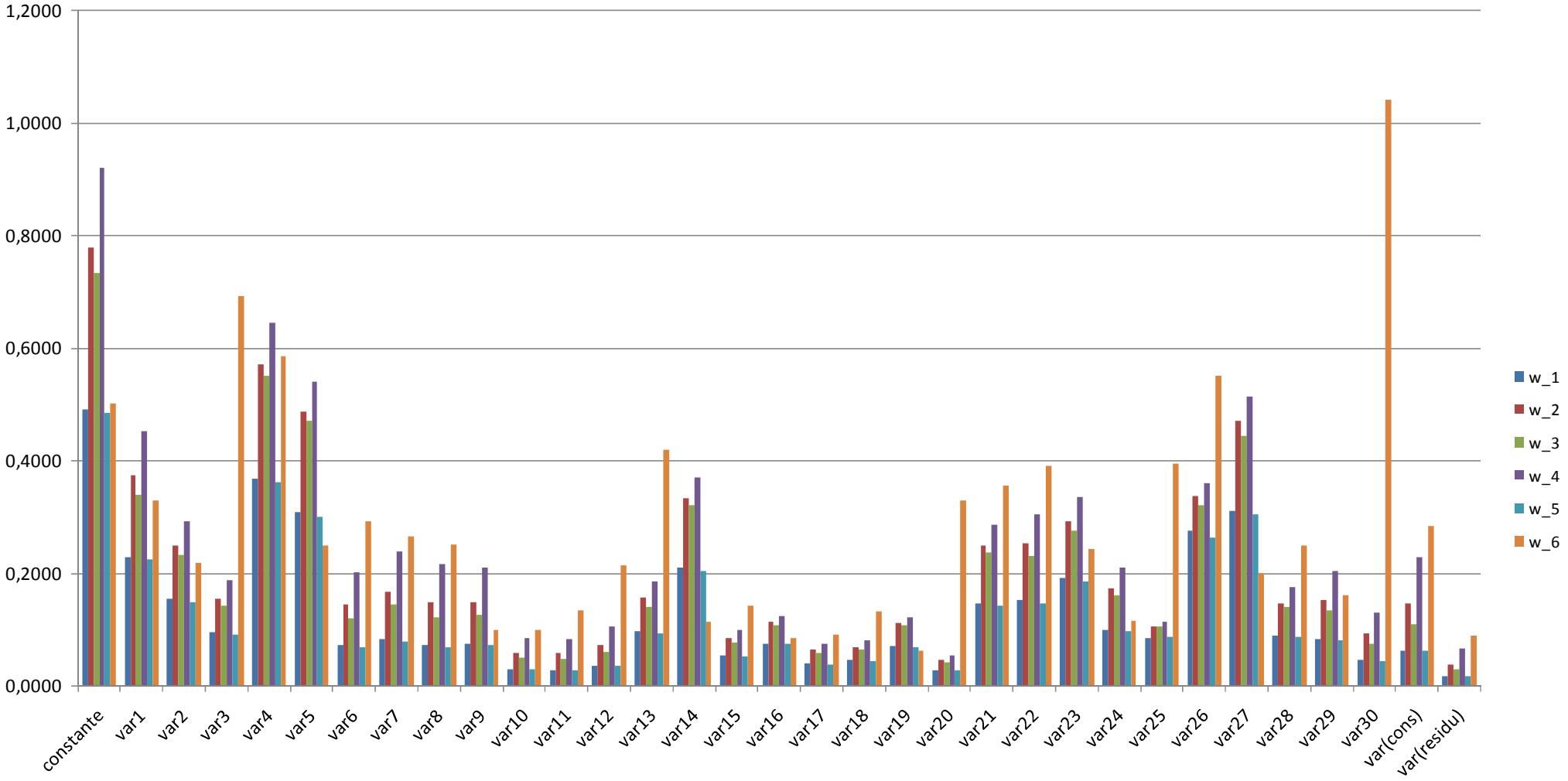
Résultat en termes de biais



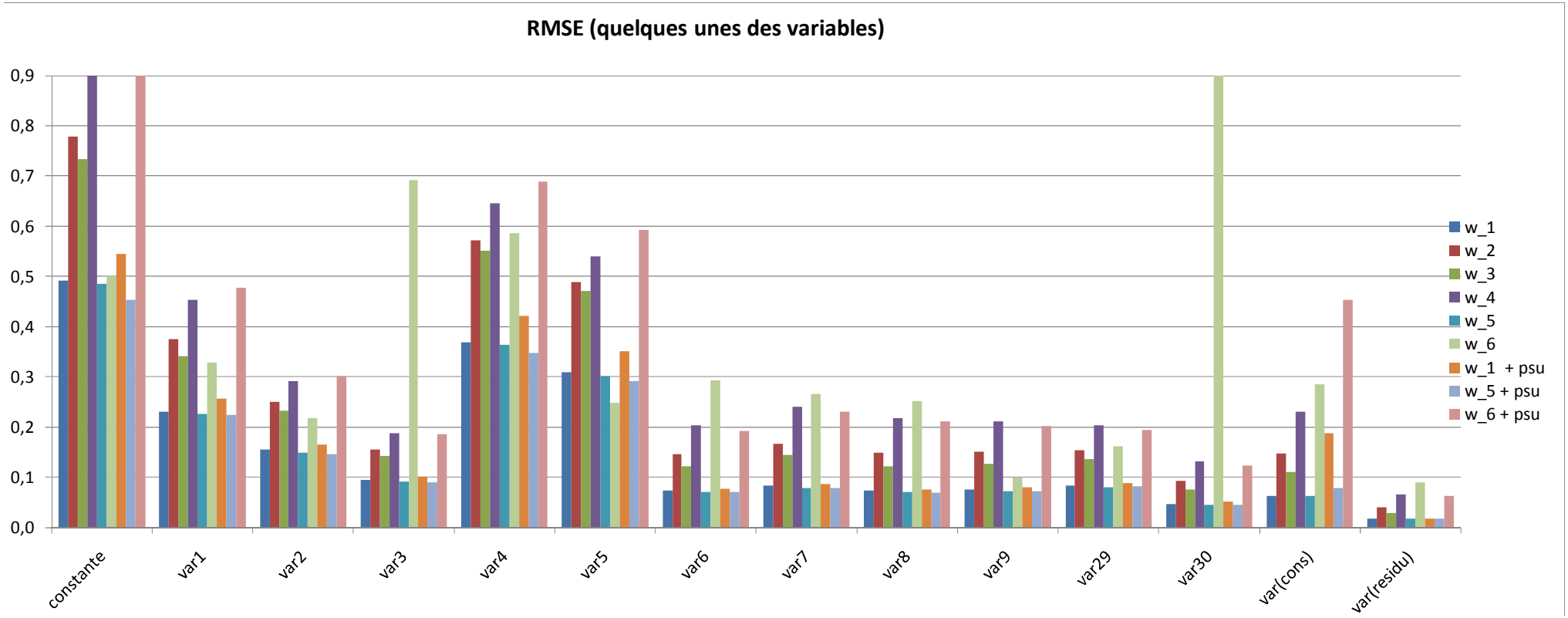
➔ On ne voit pas de « pattern » qui semble se dégager…

Résultat en termes de RMSE

Pour les 6 scénarios en modèle à 2 niveaux



Résultat en termes de RMSE



➔ 1er unweighted+psu, 2ème unweighted, 3ème w_1, 4ème w_1+psu

➔ « gap » avec les autres

Discussion

- Le « unweighted » semble être le moins mauvais…
- …en 2 niveaux ou 3 niveaux; quelques problèmes d'optimisation pour ce dernier.
- Si modèle à 2 niveaux, vouloir « rattraper » le niveau perdu par l'introduction de poids de sondage finaux pour les unités de niveau 2 n'est pas une bonne idée
- Utilisateurs de BDD publiques: attention, dans le doute ne pas utiliser de poids.

Limites et perspectives

- On n’a inspecté qu’une fenêtre de la problématique des MLM sur données d’enquête
- Cas particulier ici à cause du sous-échantillonnage; mais la population générée va nous permettre d’étudier la situation avec « 3 vrais degrés ».
- Lorsque les poids de niveau 1 différent de 1 ou basés sur proba. inégales , nous croyons qu’il est plus prudent d’utiliser des poids pour le niveau 2 (en scénario w_1) : à étudier.

Merci de votre attention

Un article relatif sera soumis prochainement dans une revue de statistiques appliquées.

Contact : J-P. Lucas, ESE/Santé/PEEI, CSTB Marne-la-Vallée
jean-paul.lucas@cstb.fr / jean-paul.lucas@etu.univ-nantes.fr