L’apport des modèles pour améliorer l’estimation sur des petites populations

Pascal Ardilly

Insee – Département des méthodes statistiques

Avril 2025

Problématique

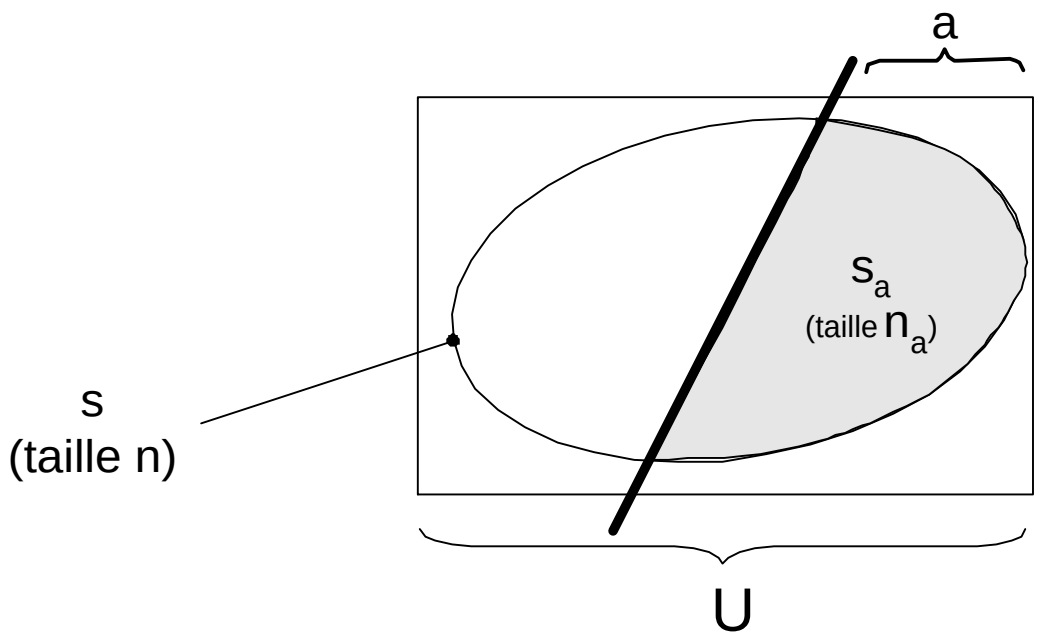
Soit une population finie U et une sous-population a.

$$Y_a = \sum_{i=1}^{N_a} Y_i \quad ?$$

Echantillon s tiré dans U (plan complexe).

π_i = probabilité de sélection de l'individu i

$s_a = s \cap a$



Négligeons l'évènement $n_a = 0$

$$\hat{Y}_a = \sum_{i \in s_a} \frac{Y_i}{\pi_i}$$

$$\Rightarrow E\hat{Y}_a = Y_a$$

Donc **pas de problème de biais** !

Mais - hélas :

$$CV\left(\hat{Y}_a\right) = O\left(\frac{1}{\sqrt{n_a}}\right)$$

Donc un sérieux problème d'instabilité si n_a petit !

QUE FAIRE ???

Objectif général : **tirer profit des corrélations entre la variable d'intérêt et des variables auxiliaires.**

Réponse 1

Un calage (si on dispose d'information auxiliaire) : **pas d'hypothèse à faire**.

Réponse 2

Exploiter l'information auxiliaire en construisant une stratégie d'estimation dépendante d'une hypothèse : **un modèle**.

Les estimateurs obtenus sont dits **indirects** : *les estimations dans un domaine donné vont explicitement dépendre des valeurs d'intérêt dans les autres domaines.*

La modélisation peut être ou non stochastique.
Si c'est le cas, il y a introduction d'une source d'aléa qui n'est plus celle de l'échantillonnage (« aléa de modèle »).

Modélisation non stochastique

I) Les estimateurs "synthétiques"

Croire à une **hypothèse descriptive de comportement en Y assimilant le domaine a au reste de la population U (le 'modèle')**;

paramètre(s) sur a = paramètre(s) sur U

puis estimer les paramètres sur **U.**

Exemple : pour estimer une moyenne $\overline{Y}_a$, postuler

$$\overline{Y}_a = \overline{Y}$$

Ici, pas d'autre aléa que celui de l'échantillonnage : l'approche reste **descriptive.**

i) Pas d'information auxiliaire

- sauf la constante ($\Leftrightarrow N_a$ connue)

$$\hat{Y}_{a,SYN} = N_a . \frac{\hat{Y}}{\hat{N}}$$

où $\hat{Y} = \sum_{i \in s} \frac{Y_i}{\Pi_i}$ et $\hat{N} = \sum_{i \in s} \frac{1}{\Pi_i}$.

L'estimateur $\hat{Y}_{a,SYN}$ mobilise tout l'échantillon s - donc des données relatives à des individus hors du domaine ($\Leftrightarrow$caractère indirect).

n grand $\Rightarrow biais \approx N_a.(\bar{Y} - \bar{Y}_a)$

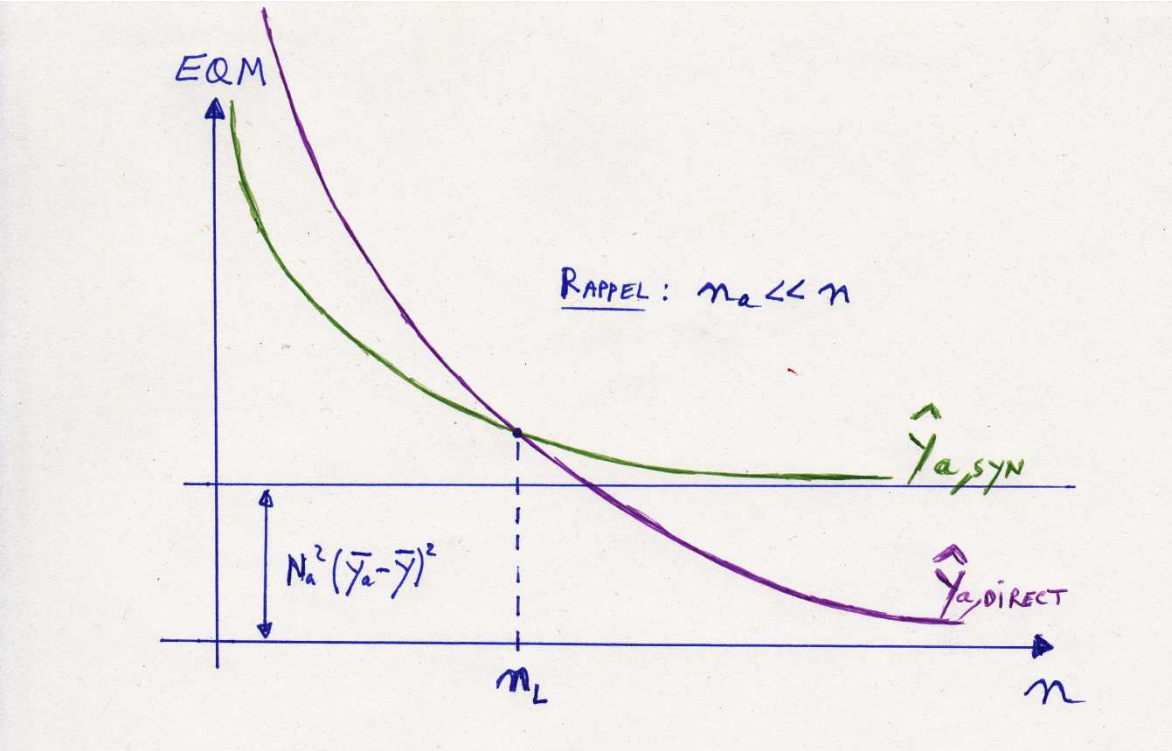
Modèle (implicite) $\bar{Y}_a = \bar{Y}$

$$EQM\left(\hat{Y}_{a,SYN}\right) = N_a{}^2.[\ \underbrace{V\left(\hat{\bar{Y}}\right)}_{\text{varie en } 1/n} + \underbrace{(\bar{Y}_a - \bar{Y})^2}_{\text{ne dépend pas de n}}\ )]$$

ERREUR FAIBLE SI LE MODELE

EST (A PEU PRES) EXACT

$\exists n_L$ tel que $n<n_L \Leftrightarrow \hat{\bar{Y}}_{a,SYN}$ préférable à $\hat{\bar{Y}}_{a,Direct}$



La problématique ressemble à celle des tirages empiriques !

ii) On dispose d'une information auxiliaire « complexe »

Dans un calage : $\hat{Y}_{calé,a} \approx X_a^T \hat{B}_a + \underbrace{\left(\hat{Y}_a - \hat{B}_a^T \cdot \hat{X}_a\right)}_{\Delta_a}$

$$\hat{B}_a = \left(\sum_{i \in s_a} \frac{X_i \cdot X_i^T}{\Pi_i} \right)^{-1} \cdot \left(\sum_{i \in s_a} \frac{X_i \cdot Y_i}{\Pi_i} \right)$$

Δ_a est « petit » devant $X_a^T \hat{B}_a$ et vaut 0 dans les cas « habituels » ($\Leftrightarrow$constante dans X).

Pour **stabiliser** $\hat{Y}_{calé,a}$ on remplace $\hat{B}_a$ par $\hat{B}$ où

$$\hat{B} = \left(\sum_{i \in s} \frac{X_i \cdot X_i^T}{\Pi_i} \right)^{-1} \cdot \left(\sum_{i \in s} \frac{X_i . Y_i}{\Pi_i} \right).$$

D'où

$$\boxed{\hat{Y}_{a,SYN} = X_a^T \hat{B}}$$

$$\hat{Y}_{a,SYN} = X_a^T \hat{B}$$

$\tilde{B}_a$ = vrai coefficient de régression dans a

$\tilde{B}$ = vrai coefficient de régression dans U

$$Biais \approx X_a^T \left(\tilde{B} - \tilde{B}_a \right)$$

Modèle (implicite) $\tilde{B}_a = \tilde{B}$

$$EQM\left(\hat{Y}_{a,REGSYN}\right) \approx \left[X_a^T \left(\tilde{B} - \tilde{B}_a \right) \right]^2 + \text{fonction de } 1/n$$

La relation entre X et Y est "universelle" mais les spécificités locales sont prises en compte par l'intermédiaire des variables auxiliaires X_i.

Une difficulté de fond

Disposer d'une l'information auxiliaire :

a) **sur chaque individu du domaine** $\left(X_a \right)$ **et dans le questionnaire** $\left(\hat{B} \right)$

b) **explicative de** Y

c) **homogène entre les deux sources**

Dans cette optique descriptive, Y peut être qualitatif – mais ce n'est pas l'approche naturelle !

Un exemple très simple

$$\forall i \in h \ : \ Y_i = \lambda_h + U_i$$

$$\hat{Y}_{a,SYN} = \sum_{h=1}^{H} N_{ah} \frac{\hat{Y}_h}{\hat{N}_h}$$

$\Rightarrow$ Estimateur (célèbre) dit '**synthétique post-stratifié'**

Modèle implicite, pour tout h :

$$\overline{Y}_{ah} = \overline{Y}_h$$

La moyenne de Y **ne dépend que de** h **et PAS du domaine.**

Une mise en œuvre facile

Si le (pseudo) modèle est linéaire multivarié, l'estimateur **synthétique** est **formellement égal** à un estimateur calé **avec la méthode linéaire** sur les totaux de chaque variable explicative (souvent des effectifs).

X_a = vecteur d'effectifs $(N_{ah})_h$ si les facteurs explicatifs sont tous qualitatifs,

$\hat{X}, \hat{Y}, \hat{B}$ = Horvitz-Thompson **avec** s **au complet**.

Si on choisit les poids initiaux $d_i\ (s) = \frac{1}{\Pi_i} \cdot \frac{\hat{N}_a}{\hat{N}}$,

$$\hat{Y}_{a,SYN} = \sum_{i \in s} d_i . Y_i + \hat{B}^T \left(\sum_{i=1}^{N_a} X_i - \sum_{i \in s} d_i . X_i \right)$$

C'est l'expression de l'estimateur issu d'un calage par la méthode **linéaire** sur les marges $\sum_{i=1}^{N_a} X_i$, **à partir du fichier complet** s et des poids initiaux $d_i\ (s)$.

EN PRATIQUE, pour obtenir n'importe quelle estimation sur le petit domaine a, **on travaille avec l'intégralité du fichier échantillon s et avec les poids calés** :

$$\hat{Y}_{a,SYN} = \sum_{k \in s} w_k^{calé} . Y_k$$

Cela évite de former des régressions pour chaque variable Y et facilite (donc) considérablement la tâche aux utilisateurs.

En revanche il faut un calage pour chaque domaine.

Application : estimation des taux régionaux de pauvreté

Domaines : les régions (métropole)

Paramètres : 6 taux de pauvreté, dont

$$\frac{\sum_{i \in s} w_i \cdot N_i \cdot 1_{R_i \leq 0.6 * R_{médian}}}{\sum_{i \in s} w_i \cdot N_i} \quad (i : \text{ménage})$$

Source enquête : SILC (enquête européenne)

*Sources calage, niveau **régional*** :

- RP : sexe / âge / diplôme / nationalité / CS / ZUS / TUU / type ménage / Locataire
- RDL : quantiles (5%) du « niveau de vie »
- CAF : effectif allocataires ASPA (personnes âgées)

Nombreux poids négatifs si la région a un comportement spécifique (Île-de-France : 1632 / 10 602 ; Corse : 1926 / 10 602).

Erreur / biais au niveau national du **nombre de pauvres**

- 0.76 % en 2009
+ 4,91 % en 2010

Grande proximité avec les « vrais » indicateurs régionaux RDL

Année 2010

REG	Taux de pauvreté monétaire Méthode directe	Taux de pauvreté monétaire Méthode « petits domaines »
11	10.73	12.43
21	14.88	14.26
22	20.09	14.05
23	13.64	12.79
24	11.15	11.58
25	9.40	13.06
26	13.86	12.37
31	18.65	18.04
41	16.99	13.52
42	12.80	11.00
43	13.94	12.66
52	9.12	11.04
53	13.53	11.05
54	14.53	13.49
72	12.86	12.50
73	14.78	13.52
74	18.40	14.16
82	9.54	11.84
83	14.05	13.62
91	18.15	17.93
93	14.56	15.25
94	28.35	18.57

Une amélioration universelle : le "Benchmarking"

Caler sur un estimateur direct $\hat{Y}_D$ **fiable** relatif à un domaine D de grande taille que les petits domaines a partitionnent :

$$\tilde{\tilde{Y}}_a = \frac{\hat{Y}_{a,SYN}}{\sum_{a \in D} \hat{Y}_{a,SYN}} . \hat{Y}_D$$

C'est en fait une **simple « règle de trois »** :

- Réduction du biais (en général)
- Diffusion locale cohérente avec la diffusion nationale.

Une faiblesse majeure : on ne peut pas (généralement) JUGER DE L'ERREUR !

$$EQM\left(\hat{Y}_{a,SYN}\right) = E\left(\hat{Y}_{a,SYN} - Y_a\right)^2$$

Car il n'est pas possible d'estimer de manière stable

(⇔faible variance) les EQM : **l'estimation de la composante de biais pose un problème de fond**.

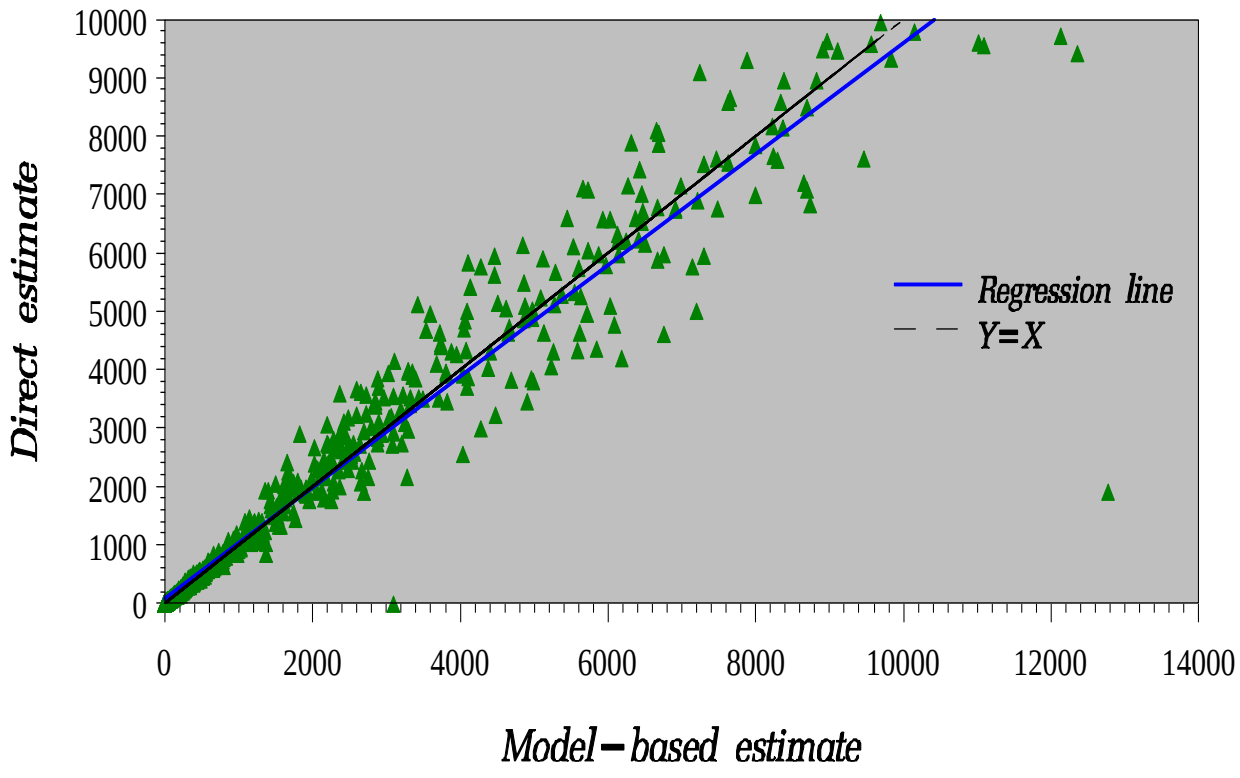
S'agissant d'erreur, **le risque vient du biais surtout** - la variance est un problème plus 'annexe'.

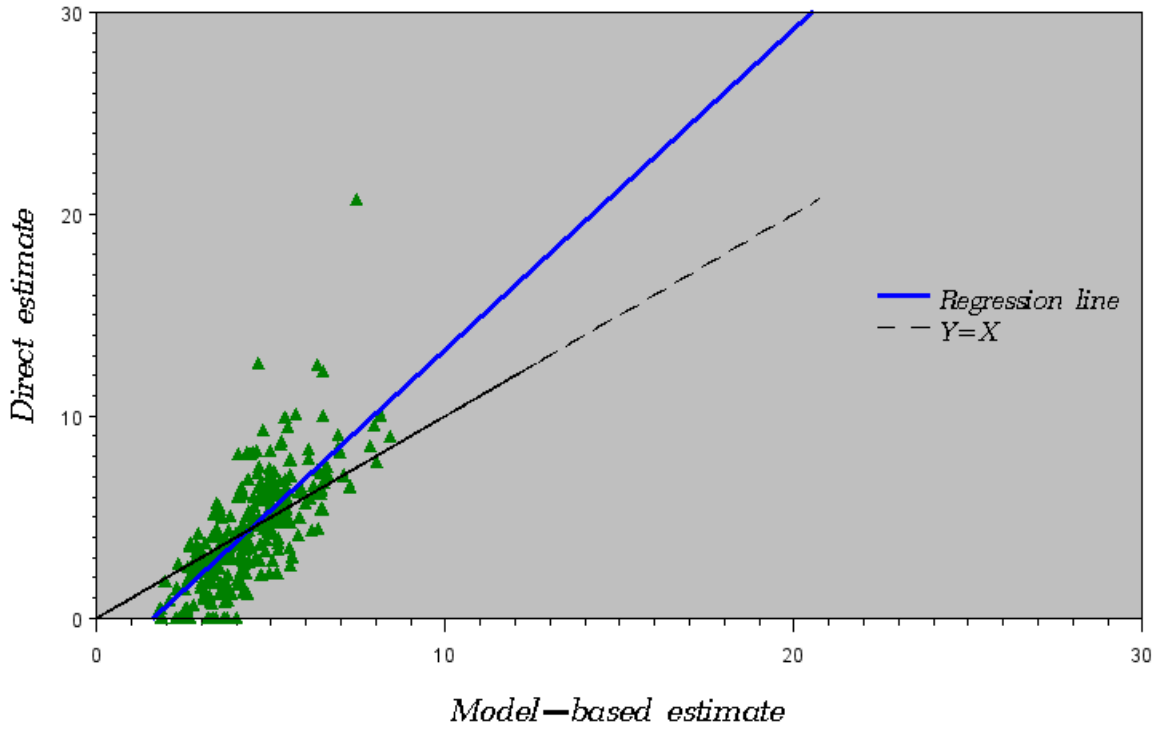
On a néanmoins deux procédures sérieuses pour détecter à un niveau GLOBAL un modèle *a priori* inadapté :

- Comparer la somme des estimations locales par modèle à l'estimation directe *nationale*;

- Régresser les estimations locales par modèle sur les estimations locales directes (approche graphique).

On positionne le nuage de points par rapport à la droite $Y = X$. On regarde sa **symétrie par rapport à cette droite**.




30
20
10
0
0
10
20
30
Direct estimate
Model−based estimate
Regression line
Y=X

Nuage dissymétrique et biais vont de pair

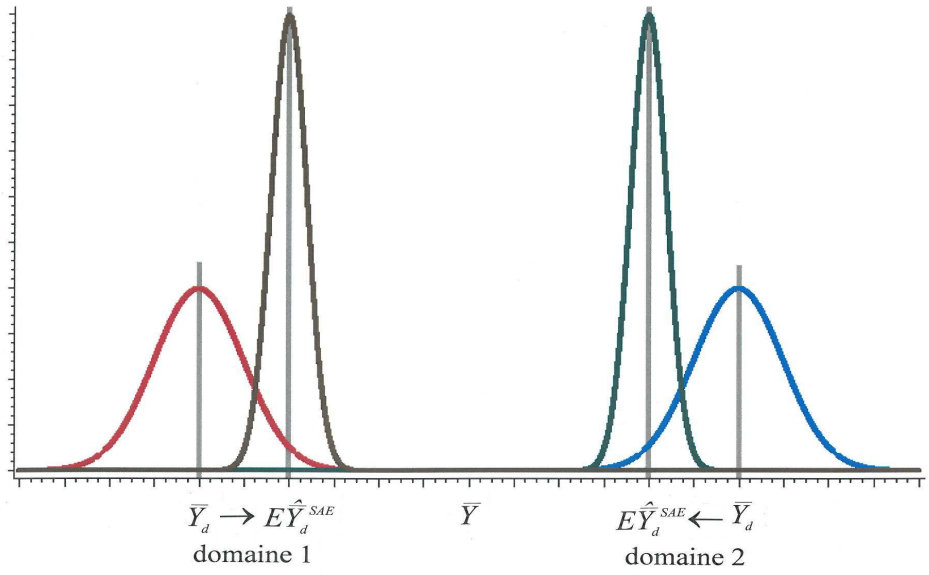
Pourquoi ce biais ?

Le biais est « naturellement » dû à un **phénomène d'uniformisation** de l'estimation 'petits domaines' consécutive à une **modélisation** : la variabilité est réduite par la forme synthétique $\overline{X}_d^T \cdot \hat{B}$ de l'estimation – plus généralement par la **contrainte générée par la modélisation** (*modèle* = *moule* !).

Dans ce cas, la distribution des $\hat{\overline{Y}}_d^{SAE}$ se « contracte » par rapport à celle des $\hat{\overline{Y}}_d^{direct}$:

- **les espérances $\hat{\overline{Y}}_d^{SAE}$ se rapprochent de $\overline{Y}$;**
- **les variances des $\hat{\overline{Y}}_d^{SAE}$ sont faibles.**

Mécanisme de biais/variance par domaine



II) Les estimateurs "composites"

$\hat{Y}_a^D$: un estimateur direct (expression quelconque)

$\hat{Y}_a^{SYN}$: un estimateur synthétique

$$\boxed{\hat{Y}_{a,COMP} = \phi_a . \hat{Y}_a^D + (1-\phi_a) . \hat{Y}_a^{SYN}} \quad \text{où} \quad \phi_a \in [0,1]$$

On module ainsi les poids des estimateurs *directs* (biais nul ou faible, forte variance) et *synthétique* (biais, faible variance).

A) Estimateur composite optimum :

On minimise l'EQM en ϕ_a (très facile) :

$$\phi_a(OPTI) = \frac{EQM\left(\hat{Y}_a^{SYN}\right)}{EQM\left(\hat{Y}_a^{SYN}\right) + EQM\left(\hat{Y}_a^{D}\right)}$$

MAIS problème bloquant d'estimation <u>stable</u> du poids.

B) Estimateur dépendant de la taille de l'échantillon :

Idée : retrouver un estimateur direct lorsque n_a est « assez grand », c'est-à-dire lorsque

$$\hat{N}_a = \sum_{i \in s_a} \frac{1}{\Pi_i}$$

est grand. Par exemple, après un choix de δ :

$$\phi_a = \begin{cases} 1 & \text{si } \hat{N}_a > \delta \cdot N_a \\ \dfrac{\hat{N}_a}{\delta . N_a} & \text{si } \hat{N}_a \leq \delta \cdot N_a \end{cases}$$

On peut espérer un biais limité, grâce à la fréquence des cas où $\phi_a = 1$.

Modélisation stochastique

On s'appuie sur un modèle utilisant des variables auxiliaires et une **composante stochastique** → rend **Y aléatoire.**

L'effet propre au domaine est isolé et apparaît **explicitement**.

L'unité modélisée peut être, selon le niveau de disponibilité des variables explicatives :

- le petit domaine **a**
- l'individu **i** de U

Différentes familles d'estimateurs, essentiellement :

- Les estimateurs sans biais linéaires optimaux ;
- Les estimateurs de comptage ;
- Les estimateurs optimaux ;
- Les estimateurs Bayésiens.

Les estimateurs sans biais optimaux linéaires (BLUP)

La variable d'intérêt Y est **quantitative** et **continue** (ou presque).

A) Cas d'une modélisation au niveau individuel (modèle "de base")

$$Y_{a,i} = X_{a,i}^T \cdot \beta + \nu_a + e_{a,i}$$

Ref : Battese, Harter, Fuller (JASA, 1988)

$X_{a,i}$ = vecteur d'effets fixes, connus partout sur le domaine - quantitatifs et / ou qualitatifs.

ν_a est un **effet aléatoire traduisant la spécificité du domaine** - au-delà des $X_{a,i}$.

$E\nu_a = 0$ ET $Var\nu_a = \sigma_v^2$ ET indépendance des ν_a

$e_{a,i}$ = terme résiduel aléatoire

$$E(e_{a,i}) = 0 \quad \text{ET} \quad V(e_{a,i}) = \sigma_e^2$$

En toute généralité, on n'a pas besoin de postuler de lois pour les aléas (… mais ça aide bien quand même !)

C'est un modèle linéaire mixte

$Var(\vec{Y}_{a,i})$ est une **matrice bloc-diagonale**.

Le paramètre à estimer $\overline{Y}_a$ (vision 'sondeur') est en fait une **variable aléatoire à prédire.**

Stratégie de prédiction / estimation ?

On cherche un prédicteur $\hat{\overline{Y}}_a^H$ **simple** :

$$\hat{\overline{Y}}_a^H = a^T \cdot \left(\vec{Y}_i\right)_{i \in s} + b$$

Vérifiant ('sans biais') $E\left(\hat{\overline{Y}}_a^H - \overline{Y}_a\right) = 0$

et minimisant l'erreur $E\left(\hat{\overline{Y}}_a^H - \overline{Y}_a\right)^2$

La solution est le **prédicteur BLUP**
(Best Linear Unbiased Predictor)

Soit $$\overline{X}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} X_{a,i}$$

On suppose qu'on connaît tous les $X_{a,i}$, donc $\overline{X}_a$.

Pratiquement, l'information individuelle $X_{a,i}$ provient du questionnaire et sa vraie moyenne $\overline{X}_a$ d'une source exhaustive (base de sondage ou pas).

RAPPEL : risque d'hétérogénéité !!!

Supposons $n_a << N_a$; l'estimateur BLUP de $\overline{Y}_a$ est :

$$\hat{\bar{Y}}_a^H = \bar{X}_a^T \tilde{\beta} + \tilde{v}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} \tilde{Y}_{a,i}$$

où $\tilde{Y}_{a,i} = X_{a,i}^t \cdot \tilde{\beta} + \tilde{v}_a$ prédicteur individuel optimum, et

$$\tilde{\beta} = \left(\sum_{a=1}^{m} \left(\sum_{i \in s_a} X_{a,i} X_{a,i}^T - \gamma_a \cdot n_a \cdot \bar{x}_a \cdot \bar{x}_a^T \right) \right)^{-1} \times \left(\sum_{a=1}^{m} \left(\sum_{i \in s_a} X_{a,i} \cdot Y_{a,i} - \gamma_a \cdot n_a \cdot \bar{x}_a \cdot \bar{y}_a \right) \right)$$

$\Rightarrow$ Expression **synthétique** impliquant TOUS les a : **il est donc (très) stable.**

et $$\tilde{v}_a = \gamma_a \cdot \left(\bar{y}_a - \bar{x}_a^T \cdot \tilde{\beta} \right)$$

$$\bar{x}_a = \frac{1}{n_a} \sum_{i \in s_a} X_{a,i} \qquad \bar{y}_a = \frac{1}{n_a} \sum_{i \in s_a} Y_{a,i}$$

$$\gamma_a = \frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_e^2}{n_a}}$$

On a aussi

$$\hat{\bar{Y}}_a^H = \gamma_a \cdot \left[\bar{y}_a + \left(\bar{X}_a - \bar{x}_a \right)^T \tilde{\beta} \right] + \left(1 - \gamma_a \right) \cdot \bar{X}_a^T \tilde{\beta}$$

Partie (pseudo) directe Partie synthétique

Si $n_a = 0$ on **convient** de retenir $\hat{\bar{Y}}_a^H = \bar{X}_a^T \tilde{\beta}$.

Packages R : Sae, Emdi

$\hat{\bar{Y}}_a^H$ est bien une statistique **linéaire** en $Y_{a,i}$.

σ_v^2 petit $\Rightarrow$ peu d'effet propre au modèle $\Rightarrow$ modèle bien spécifié $\Rightarrow$ partie synthétique prime

n_a grand $\Rightarrow$ partie directe prime

Considérant seulement l'aléa de sondage, cet estimateur n'a pas de bonnes propriétés (biais, non convergent) : normal, **les poids de sondage n'interviennent pas !**

C'est bien totalement **dépendant du modèle.**

Il faut *in fine* estimer σ_e^2 et σ_v^2 et aboutir à un estimateur empirique dit EBLUP (E=*empirical*) - qui est le (seul) "véritable estimateur".

Sous hypothèse de loi de Gauss, estimation par MV (processus itératif de Newton, Scoring de Fisher,…),

$\Longrightarrow$ d'où $\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ puis $\hat{\gamma}_a$ et $\hat{\beta}$, et *in fine* $\hat{\bar{Y}}^{H}_{a,EBLUP}$

Méthode alternative : méthode des moments - plus simple, mais moins efficace si $Y \approx$ gaussien.

On a le droit de mal spécifier le modèle utilisé pour l'inférence MAIS les variables explicatives 'oubliées' ne doivent pas impacter la composition de l'échantillon.

$\Rightarrow$ **nécessité d'une hypothèse d'échantillonnage non informatif.**

Techniquement, il faut et il suffit :

$\tilde{\vec{X}}_i$ = variables utilisées pour tirer S (ex : critère de taille)

$$Loi(\vec{Y}_i \,|\, \vec{X}_i \,, \tilde{\vec{X}}_i \,) = Loi(\vec{Y}_i \,|\, \vec{X}_i \,)$$

Solution **:** inclure $\tilde{X}_a$ dans le vecteur des régresseurs X_a

Propriété **toujours vraie avec un sondage aléatoire simple** (puisque $\tilde{X} = 1$).

Calcul d'erreur par méthode analytique

$$E\left(\hat{\bar{Y}}_a^H - \bar{Y}_a\right)^2 = \gamma_a \cdot \frac{\sigma_e^2}{n_a} + O\left(\frac{1}{n}\right)$$

Estimation de l'erreur de l'EBLUP **(opérationnel !**)

$$\hat{E}\left(\hat{\bar{Y}}_a^{EBLUP} - \bar{Y}_a\right)^2 = \hat{\gamma}_a \cdot \frac{\hat{\sigma}_e^2}{n_a} + g(\hat{\sigma}_v, \hat{\sigma}_e^2) + o\left(\frac{1}{n}\right)$$

L'estimation naïve

$$\hat{\gamma}_a \cdot \frac{\hat{\sigma}_e^2}{n_a}$$

est acceptable mais délaisse les termes en $\frac{1}{n}$.

ALTERNATIVE : *méthodes Jackknife et Bootstrap*.

 Les erreurs restent en $\frac{1}{n_a}$: ce sont donc (seulement) γ_a et σ_e^2 qui « font la différence » avec l'estimation directe.

C'est pour cela qu'*in fine la précision peut apparaitre médiocre si on conserve les standards de la qualité nationale.*

Si σ_v^2 et σ_e^2 petits (modèle bien adapté) $\Rightarrow$ fort gain

Cas d'une modélisation au niveau du domaine, prise en compte du plan de sondage (modèle de Fay et Herriot)

Ref : Fay, Herriot (JASA, 1979)

$$g(\bar{Y}_a) = \bar{X}_a^T \cdot \beta + b_a \cdot v_a$$

$\beta \in R^p$, $b_a \in R$ **connu**

v_a variable aléatoire (« effet aléatoire » propre au domaine)

$$E(v_a) = 0 \text{ et } V(v_a) = \sigma_v^2$$

Les v_a sont mutuellement indépendants

- pas d'hypothèse de loi de v_a (mais ça aide bien…)
- le modèle porte sur la vraie valeur $\bar{Y}_a$.

Cela suppose que $\bar{Y}_a$ *est quantitative et de nature continue*. On l'acceptera pour des proportions P_a ou des dénombrements si N_a est 'grand'.

$$g(\hat{\bar{Y}}_a) = g(\bar{Y}_a) + e_a$$

e_a = erreur d'échantillonnage, *supposée sans biais*, de (vraie) variance Ψ_a (en pratique variance **estimée)**. Les e_a sont supposés mutuellement indépendants.

$$\boxed{g(\hat{\bar{Y}}_a) = \bar{X}_a^T \beta + b_a v_a + e_a}$$

On considérera v_a et e_a comme indépendantes.

C'est un modèle linéaire mixte,

qui mêle 2 natures d'aléas

Le paramètre à estimer est (β, σ_v^2).

La modélisation considérant v_a comme effet FIXE est plus naturelle mais le modèle n'est plus estimable : l'effet local se justifie surtout par la recherche de « parcimonie ».

Si spécificité locale connue (ex : fermeture de société - pour expliquer le chômage dans une zone) $\Longrightarrow$ prise en compte par une indicatrice dans $\bar{X}_a$, en sus des v_a - pour tenter de réduire σ_v^2.

$\overline{Y}_a$ est la **variable aléatoire** à prédire.

Si $\overline{Y}_a$ est une proportion $\mathrm{P_a}$, prendre $g(x) = Log\dfrac{x}{1-x}$ (ainsi $g(P_a)$ décrit $\mathbb{R}$) ou encore $\mathrm{g(x) = \arcsin\sqrt{x}}$ (les variances d'échantillonnage ne dépendent plus des $\mathrm{P_a}$).

Stratégie BLUP, dans le cas $g(x) = x$:

$$\boxed{\hat{\overline{Y}}_a^H = \overline{X}_a^T \tilde{\beta} + \gamma_a \cdot \left(\hat{\overline{Y}}_a - \overline{X}_a^T \tilde{\beta}\right)}$$

ou encore $$\hat{\overline{Y}}_a^H = \gamma_a \cdot \underbrace{\hat{\overline{Y}}_a}_{\text{Estimateur direct}} + (1-\gamma_a) \cdot \underbrace{\overline{X}_a^T \tilde{\beta}}_{\text{Estimateur synthétique}}$$

$$\tilde{\beta} = \left[\sum_{a=1}^{m} \frac{\overline{X}_a \cdot \overline{X}_a^T}{\Psi_a + b_a^2 \cdot \sigma_v^2}\right]^{-1} \cdot \left[\sum_{a=1}^{m} \frac{\overline{X}_a \cdot \hat{\overline{Y}}_a}{\Psi_a + b_a^2 \cdot \sigma_v^2}\right]$$

et $$\gamma_a = \frac{b_a^2 \cdot \sigma_v^2}{\Psi_a + b_a^2 \cdot \sigma_v^2} = \frac{Variance\ stochastique}{Variance\ totale}$$

On a donc γ_a = part de la variance totale due au modèle.

$\tilde{\beta}$ est stabilisé par la présence de m domaines : on a bien $V(\tilde{\beta})=\mathrm{O}\left(\frac{1}{\mathrm{n}}\right)$.

* Si σ_v^2 est petit $\Rightarrow$ l'influence de v_a est faible $\Rightarrow$ le modèle est efficace $\Rightarrow$ $\hat{\bar{Y}}_a^H$ est « presque » l'estimateur synthétique.

* Si Ψ_a est faible, γ_a tend vers 1 et l'estimateur direct reprend l'avantage.

Si $n_a = 0$, on convient que $\hat{\bar{Y}}_a^H = \bar{X}_a^T \tilde{\beta}$.

Avec seulement l'aléa de sondage, cet estimateur
- est convergent ($n_a \to N_a$) puisque $\gamma_a \to 1$
- est biaisé !

En pratique, instabilité de l'estimation de variance **d'échantillonnage** locale : il faut lisser les variances d'échantillonnage $\hat{\Psi}_a$ sinon risque d'obtenir $\hat{\sigma}_v^2 = 0$.

Le biais est nul seulement si on considère les 2 aléas.

Erreur totale / cas où m **est grand)**

$$EQM\left(\hat{\bar{Y}}_a^H\right) \approx \gamma_a \Psi_a + O\left(\frac{1}{m}\right) \approx \gamma_a \Psi_a$$

$$\Rightarrow \quad \boxed{\frac{EQM(\hat{\bar{Y}}_a^H)}{EQM(\hat{\bar{Y}}_a)} \approx \gamma_a}$$

Conclusion : **si** γ_a **petit (m grand)** $\Rightarrow$ **gain important**.

Comme σ_v^2 **est inconnu**, il faut l'estimer (EBLUP).

Différentes méthodes (maximum de vraisemblance…)

$$\hat{E}\left(\hat{\bar{Y}}_a^{EBLUP} - \bar{Y}_a\right)^2 = \hat{\gamma}_a \cdot \hat{\psi}_a + g(\hat{\sigma}_v^2) + o\left(\frac{1}{m}\right)$$

ALTERNATIVE : *méthodes Jackknife et Bootstrap* (mais très compliqué – multitude d'étapes à enchainer).

Les estimateurs SBOL et ESBOL ont le mérite d'associer « harmonieusement » les deux grandes approches de l'estimation / prédiction :

- l'approche sondage (pas de dépendance envers un modèle de comportement)
- l'approche classique par modélisation (le modèle de comportement est déterminant).

en donnant priorité à celle qui semble la plus fiable.

Fay et Herriot recommandent d'utiliser plutôt :

$$\hat{\bar{Y}}_{a,H}^{*} = \begin{cases} \hat{\bar{Y}}_a^{H} & SI \left| \hat{\bar{Y}}_a^{H} - \hat{\bar{Y}}_a \right| \leq c.\sqrt{\Psi_a} \\ \hat{\bar{Y}}_a - c.\sqrt{\Psi_a} & SI\ \hat{\bar{Y}}_a^{H} < \hat{\bar{Y}}_a - c.\sqrt{\Psi_a} \\ \hat{\bar{Y}}_a + c.\sqrt{\Psi_a} & SI\ \hat{\bar{Y}}_a^{H} > \hat{\bar{Y}}_a + c.\sqrt{\Psi_a} \end{cases}$$

Ca ressemble à un test d'appartenance de $\hat{\bar{Y}}_a^{H}$ à un IC construit autour de $\hat{\bar{Y}}_a$.

La troncature va réduire de fait la variance.

On peut étendre toute cette méthodologie à des modèles plus sophistiqués, par exemple :

- *Modèles de corrélation spatiale :*

$$Cov(v_a, v_b) = \alpha \cdot e^{-\beta . d_{ab}} \qquad (\alpha, \beta) \in \mathbf{R}^2$$

- *Modèles temporels* (modèle de Rao et Yu)

$$\begin{cases} \hat{\theta}_{at} = \theta_{at} + e_{at} \\ \theta_{at} = g(\bar{Y}_{at}) = Z_{at}^T \beta + b_a v_a + u_{at} \end{cases}$$

avec $u_{at} = \rho \cdot u_{a,t-1} + \varepsilon_{at}$

→ paramètre supplémentaire = $(\rho, \sigma_\varepsilon^2)$.

→ package R : SAERY

L'objectif est d'augmenter le nombre d'observations en restreignant l'augmentation du nombre de paramètres : on gagne beaucoup en stabilité (mais risque plus grand de mauvaise spécification du modèle).

Application : estimation du nombre de chômeurs par Zone d'Emploi

$\hat{P}$ = rapport de l'estimateur pondéré du nombre de chômeurs à l'estimateur pondéré de la taille de ZE (> 14 ans)

ou

$$\hat{N}_{cho} = N_{ZE} \cdot \hat{P}$$

France (métropole) : $\hat{P} = \frac{2\,416\,400}{49\,745\,000} \approx 4{,}86\%$

Modèle définitif à 5 variables (+ la constante)

Variable explicative	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-4.70	1.64	134.7	**-2.86**	0.005
Recherche_emploi	0.65	0.25	223.2	**2.64**	0.009
En_couple	0.12	0.049	126.3	**2.56**	0.012
H_diplomés_15_18ans	111.22	50.77	207.1	**2.19**	0.030
H_non_diplomés_50_64ans	3.09	1.85	217.7	**1.67**	0.096
Total_30_49_ans	-3.27	2.76	190.9	**-1.19**	0.237

$$\hat{\psi}_a = deff_a \times \frac{\tilde{p}_a(1-\tilde{p}_a)}{n_a}$$

- $\mathrm{deff}_a = \dfrac{\mathrm{Var}(\hat{\mathrm{P}}_a)}{\mathrm{Var}_{\mathrm{sas}}(\hat{\mathrm{P}}_a)}$ (*design effect*) : obtenu à partir d'une enquête Emploi, mais au niveau régional ;

- $\tilde{\mathrm{p}}_a$ = estimateur synthétique de P_a, produisant la statistique officielle ;

Dans ces conditions, on vérifie :

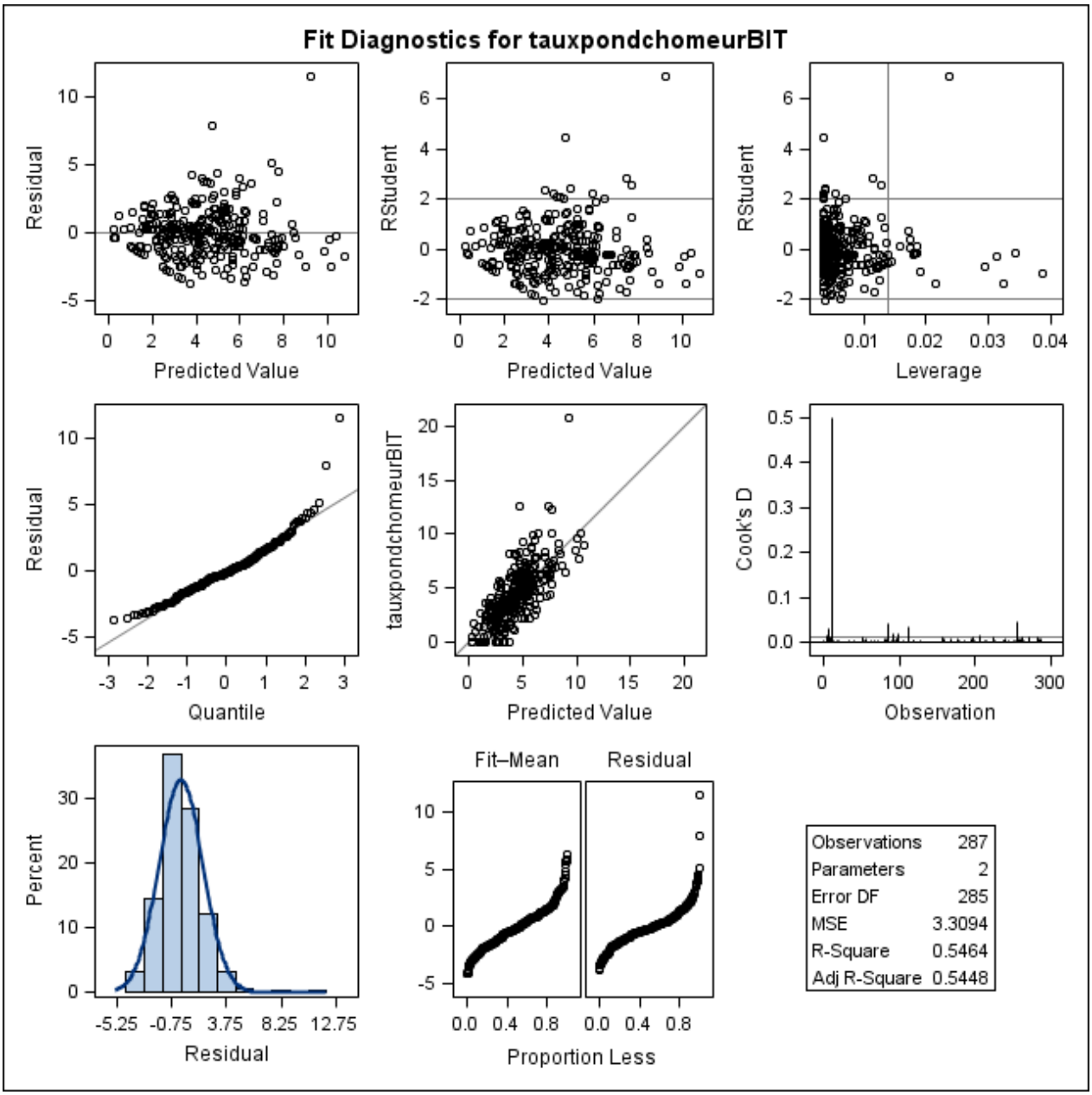
11,2% < CV des $\hat{P}_a$ < 616,6 %,

avec $Q1 = 33{,}8\%$, médiane = $51{,}1\%$ et $Q3 = 72{,}5\%$

Distribution des estimateurs
(filtre : $n_a \geq 50$ répondants) :

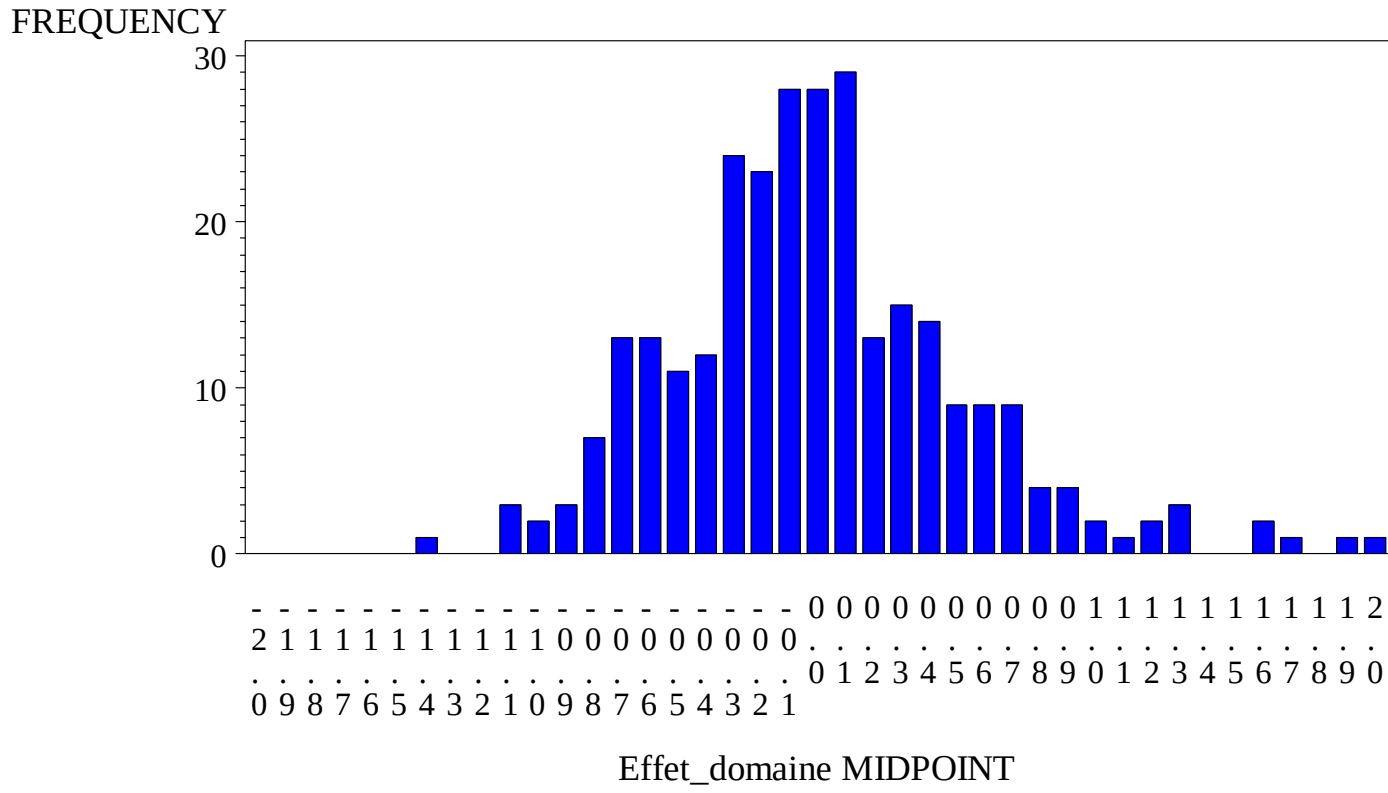
Quantile		Fay & Herriot (%)	Direct $\hat{P}_a$ (%)
100%	Max	8.40	20.7
99%		8.02	14.2
95%		6.51	10.0
90%		6.16	8.2
75%	Q3	5.21	6.2
50%	Median	4.34	4.2
25%	Q1	3.47	2.4
10%		2.85	0.0
5%		2.58	0.0
1%		1.86	0.0
0%	Min	1.82	0.0

$\hat{\sigma}_v^2 = 1{,}111$ avec un $CV \approx 30\%$


Fit Diagnostics for tauxpondchomeurBIT
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
tauxpondchomeurBIT
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 287
Parameters 2
Error DF 285
MSE 3.3094
R-Square 0.5464
Adj R-Square 0.5448

Distribution des effets domaines prédits $\hat{v}_a$

Les effets locaux $\hat{v}_a$ par ZE se répartissent entre -1,5 et 2 points de pourcentage.



Distribution du coefficient $\hat{\gamma}_a$ *modèle définitif*

Quantile		Gamma
100%	Max	0.74
99%		0.71
95%		0.57
90%		**0.49**
75%	Q3	0.33
50%	Median	**0.21**
25%	Q1	0.14
10%		0.10
5%		0.09
1%		0.07
0%	Min	0.05

Dans un peu plus de 90% des cas, priorité est donnée à l'estimateur synthétique.

Dans la moitié des ZE, l'estimateur direct issu de l'enquête Emploi contribue à plus de 20% dans la valeur de l'estimation composite finale de Fay et Herriot - ce qui est loin d'être négligeable.

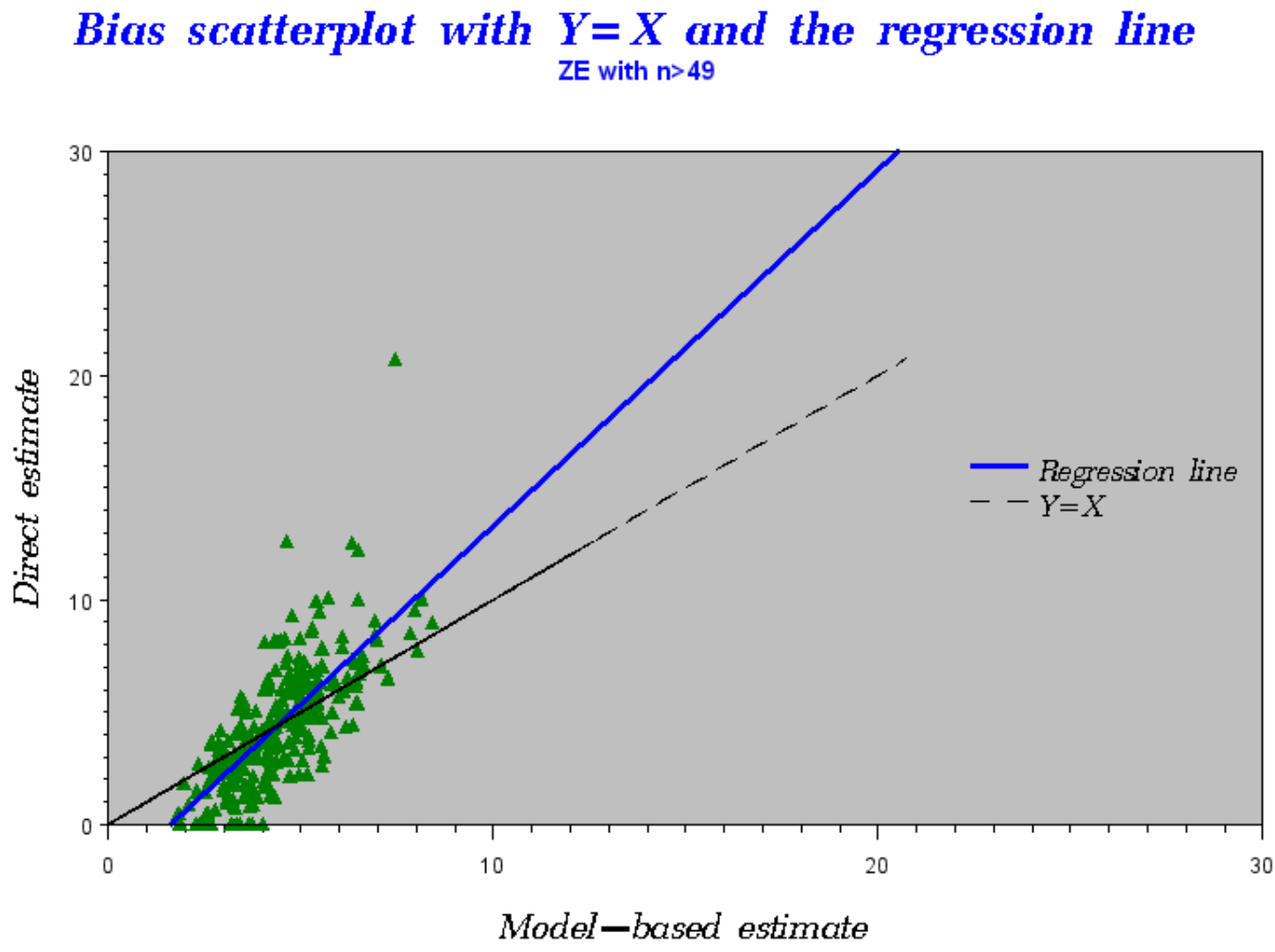
On compare l'estimation (pseudo) nationale FH et l'estimation (pseudo) nationale directe issue de l'enquête Emploi, soit

$$\sum_{ZE} N_{ZE} \cdot \frac{\hat{Y}_{ZE}}{\hat{N}_{ZE}}$$

Sur les 287 ZE concernées ($n_a \geq 50$ répondants)

Estimateur (pseudo) national	Estimation totale
Fay et Herriot	2 339 000
Enquête Emploi (*direct*)	**2 340 000**
Méthodologie actuelle Insee *(estimateur de type synthétique)*	2 305 000

Nota : la proximité entre l'estimation F&H et l'estimation directe est ici *exceptionnellement* bonne.



S**hrinkage significatif = risque de biais** pour l'aléa d'échantillonnage
(ce qui modére le résultat très satisfaisant obtenu ci-dessus…).

Les estimateurs de comptage

Un exemple : la modélisation Logistique au niveau de l'individu

Paramètre : taille d'une population c dans a, notée N_a^c

P_a = la proportion associée.

On définit $Y_{a,i} = 1$ si $i \in c$ et 0 sinon.

$$P_a = \bar{Y}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} Y_{a,i} = \frac{N_a^c}{N_a}$$

Modèle de Bernoulli : $Y_{a,i} \rightarrow B(1, P_{a,i})$

$$\forall a, \forall i \quad Log \frac{P_{a,i}}{1 - P_{a,i}} = X_{a,i}^T \beta + \nu_a$$

$$\nu_a \rightarrow N(0, \sigma_\nu^2)$$

Les $P_{a,i}$ sont des **variables aléatoires** - à prédire.

C'est un **modèle linéaire mixte généralisé**.

Il n'y a jamais d'aléa d'échantillonnage dans ce contexte, mais un empilement de 2 aléas stochastiques : aléa portant sur $P_{a,i}$ puis aléa de génération des $Y_{a,i}$.

Estimations $\hat{\beta}$ et $\hat{\sigma}_v^2$ obtenues pas EMV – mais pas directement (densité trop complexe) : on passe par un modèle **linéaire** mixte approché ou des méthodes d'approximation d'intégrales.

On obtient par ces EMV : $\hat{\nu}_a$ (théorie BLUP), puis *in fine*

$$\hat{N}_a^{c,H} = \sum_{\substack{i \in s \\ i \in a}} 1_{i \in c} + \sum_{\substack{i \notin s \\ i \in a}} \hat{E} 1_{i \in c}$$

$$\hat{N}_a^{c,H} = \sum_{\substack{i \in s \\ i \in a}} 1_{i \in c} + \sum_{\substack{i \notin s \\ i \in a}} \frac{\exp\left(X_{a,i}^T \cdot \hat{\beta} + \hat{\nu}_a\right)}{1 + \exp\left(X_{a,i}^T \cdot \hat{\beta} + \hat{\nu}_a\right)}$$

On obtient donc des estimateurs de nature synthétique et cela va limiter leur variance.

Si on se limite à l'aléa de sondage - qui existe puisqu'il y a un échantillon - de forts biais sont possibles.

Code R disponible : voir package ***GLMM*** ou ***lme4***.

Estimation du nombre de chômeurs par Zone d'Emploi

Choix des variables auxiliaires

Le choix est fort restreint par 2 conditions :

1) Une information $X_{a,i}$ présente dans l'enquête Emploi (pour l'ajustement du modèle);

2) Calcul des $\hat{P}_{a,i} \Rightarrow$  les valeurs $X_{a,i}$ doivent absolument être disponibles **au niveau individu** sur une base **exhaustive** (ou extrapolable - comme le RP).

Seule base adaptée : le recensement !

On est donc très contraint, car le recensement est limité en variables potentiellement explicatives ET présentes dans l'EE.

En pratique, on doit hélas se limiter à :

- La situation déclarée pour le mois en cours (déclaration spontanée de l'état de chômage ou non)

- La recherche ou non d'un emploi

- Le sexe
- L'âge
- La nationalité
- Le diplôme le plus élevé
- L'indicateur de vie en couple ou non
- Le statut matrimonial
- Le statut d'occupation du logement

Les 2 variables les plus prometteuses *a priori* (en vert) conduisent à des estimations locales inacceptables (sommation nationale largement excessive).

$\Rightarrow$ c'est le problème **d'hétérogénéité** des informations explicatives.

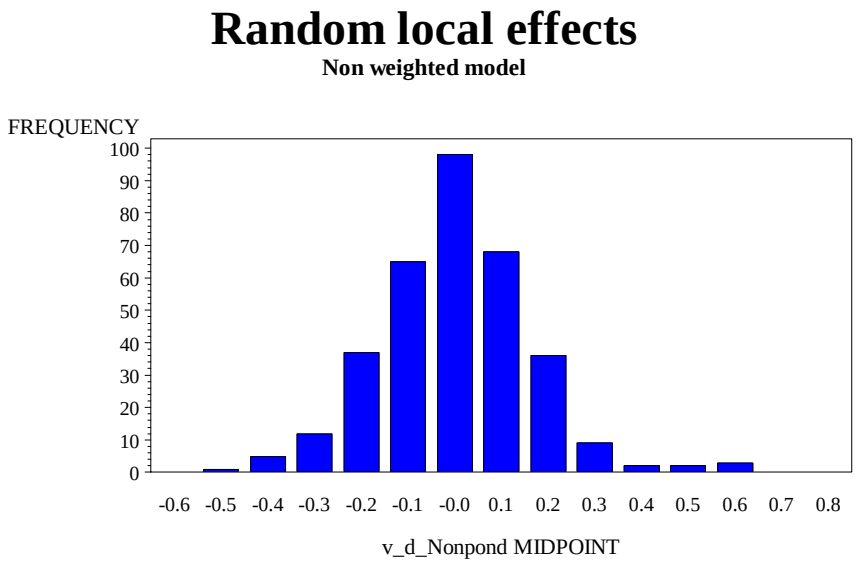
Conclusion : on devra s'en passer !

Modèle finalement retenu (un peu décevant…) :

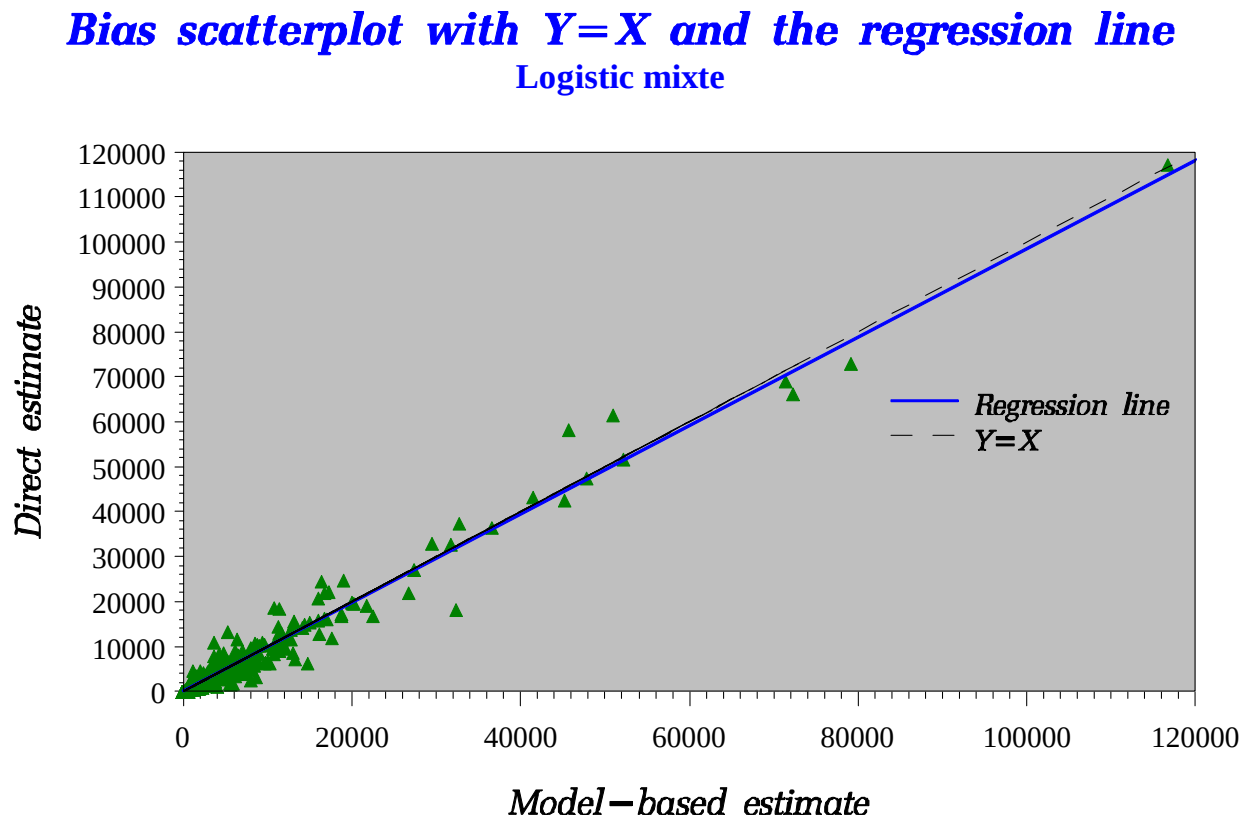
Effect	Estimateur	Rappel estima^{t} NON pondéré	Std	t Value	Pr > |t|
Intercept	-6.76	-6.15	0.1215	-55.62	<.0001
stoc_loc_1	-0.69	-0.76	0.00155	-444.74	<.0001
stoc_loc_2	0	0	.	.	.
AGE_1	3.31	3.49	0.01429	231.48	<.0001
AGE_2	5.10	5.33	0.01403	363.31	<.0001
AGE_3	4.99	5.26	0.01403	355.68	<.0001
AGE_4	4.80	5.00	0.01389	345.50	<.0001
AGE_5	4.40	4.52	0.01391	316.02	<.0001
AGE_6	0	0	.	.	.
dipl_bin_1	-0.50	-0.53	0.00153	-324.93	<.0001
dipl_bin_2	0	0	.	.	.
Nat_1	-0.53	-0.54	0.00243	-219.97	<.0001
Nat_2	0.12	0.11	0.00358	34.50	<.0001
Nat_3	0	0	.	.	.
matr_1	-0.58	-0.55	0.00171	-338.37	<.0001
matr_2	0	0	.	.	.

Distribution des *effets locaux* $\hat{v}_a$ (*approche NON pondérée*)

$$\hat{\sigma}_v^2 = 0.082 \quad \text{(écart-type = 0.017)}$$



Pas de shrinkage significatif : *a priori* **absence de biais**.



Estimations par ZEAT avant benchmarking

Comparaison (relative) à l'estimation DIRECTE

ZEAT	Classique ($\sigma_v^2 = 0$)	Mixte ($\sigma_v^2 > 0$)	Insee (officiel*)
1	+ 8.1 %	+ 0.7 %	- 10.6 %
2	- 0.4 %	+ 0.9 %	+ 2.4 %
3	- 29.1 %	- 0.5 %	- 8.2 %
4	+ 11.8 %	- 3.0 %	+ 10.6 %
5	+ 8.6 %	+ 0.5 %	+ 8.1 %
7	- 2.6 %	- 6.4 %	+ 0.4 %
8	+ 16.6 %	+ 2.8 %	+ 3.5 %
9	- 16.8 %	- 0.1 %	+ 1.2 %
TOTAL	**Sigma = 94,0**	**Sigma = 14,9**	**Sigma = 45,0**

(*) Méthode *ad hoc* de type synthétique

Donc **l'intégration d'effets locaux est très efficace**.

Estimations nationales (avant benchmarking) :

Direct :	**2 416 000**	(= référence)
Classique :	2 409 000	
Mixte :	2 409 000	
Officiel :	2 408 000	

Quelques questions de fond

1) Sur le principe, accepte-t-on les estimations dépendantes de modèles (nota : on le fait pour la non-réponse !) - et donc le biais ?

Modèle = hypothèse simplificatrice de la réalité

Rôle et danger des effets locaux

La présence de termes aléatoires « locaux » v_d constitue un échappatoire confortable

$$Y_{d,i} = X_{d,i}^T \cdot B + v_d + e_{d,i} \quad \text{et} \quad Var(v_d) = \sigma_v^2 > 0$$

mais on peut « payer » cette facilité par un σ_v^2 (*bien trop*) *grand.*

2) Que penser de la pertinence d'un modèle sur les parties des domaines qui ne sont pas observées ?

$\rightarrow$ sur l'échantillon, on a des outils de diagnostic (indicateurs, tests, graphiques…).

Mais une spécificité de Y dans d peut très bien passer inaperçue : un bon ajustement global n'est pas une garantie de pertinence de toute estimation locale.

3) Les modèles stochastiques sont-ils meilleurs *(versus estimateur « descriptif » = estimateur synthétique) ?*

- n'apportent rien si les domaines sont de très petite taille ($\gamma_d \approx 0$: on récupère du synthétique $\approx$ pur).

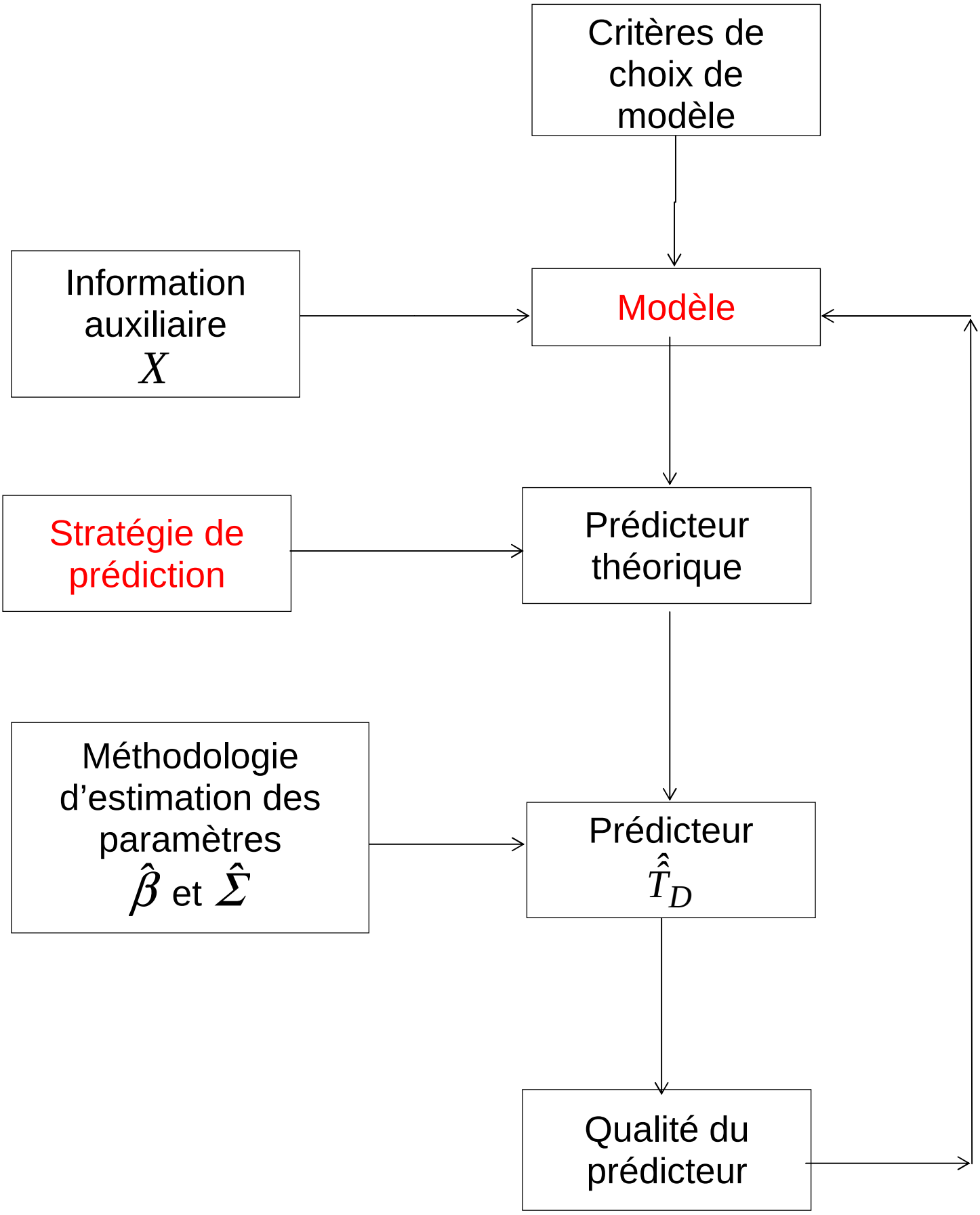
- permettent quand même en général de réduire les biais **de sondage** des estimateurs - via l'introduction de v_d si toutefois σ_v^2 ***reste modéré.***

4) Modèle niveau domaine ou niveau individu ?

CONCLUSION

On travaille donc avec des **présomptions** : le modèle constitue toujours un **acte de foi** - comme dans tout contexte de prévision !

*** Choix de méthode & modèle ***



5) Faut-il accepter que les poids de sondage ne soient pas impliqués ?

→ **l'implication des poids reste possible**:

Exemple : Estimateurs pseudo-EBLUP (convergent)

$$\overline{X}_d^T \cdot \hat{B}_w + \hat{\gamma}_{d,w} \cdot \left(\hat{\overline{Y}}_{d,w} - \hat{\overline{X}}_{d,w}^T \cdot \hat{B}_w \right)$$

6) Le concept d'erreur est-il le bon ?

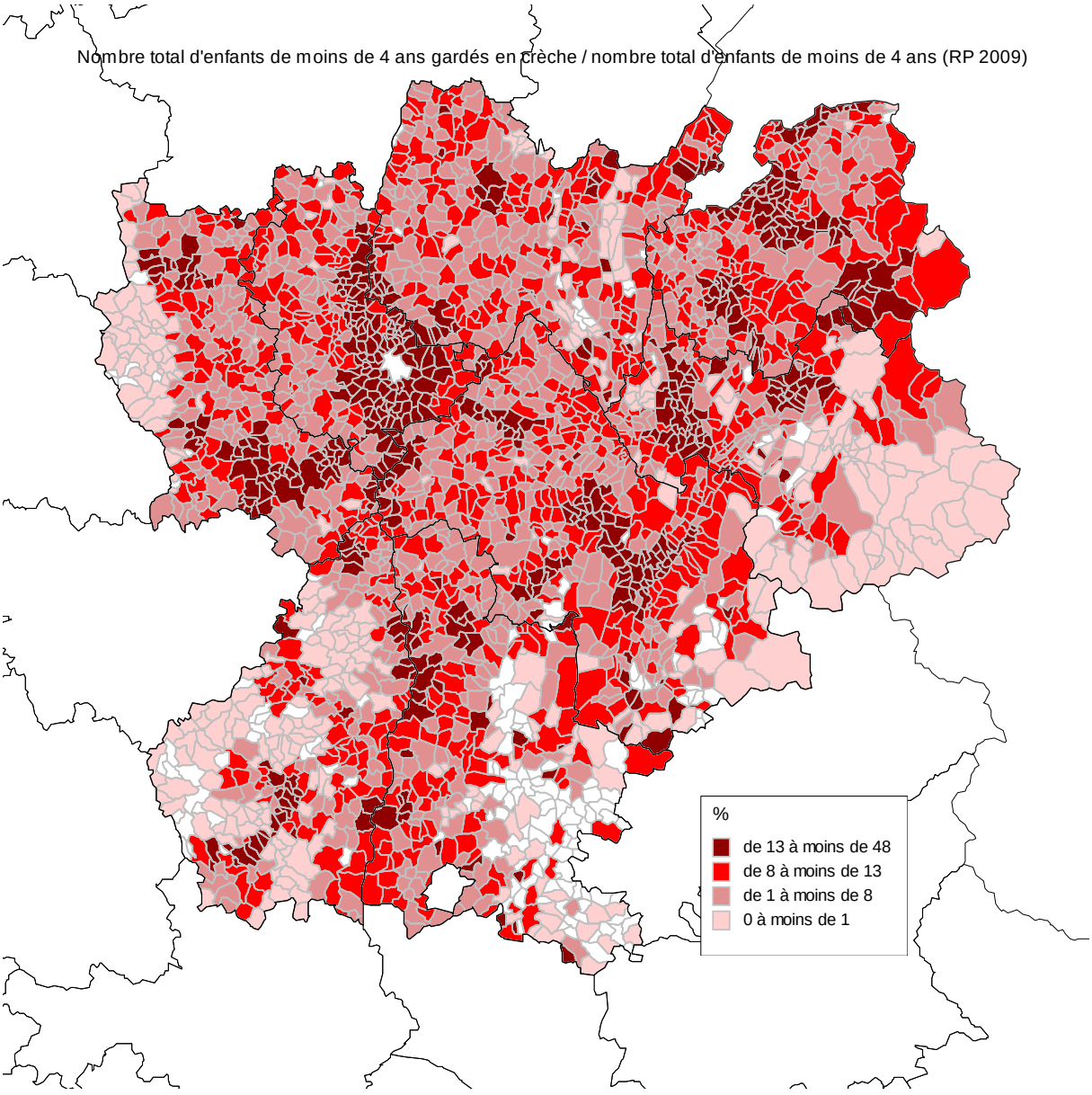
ATTENTION : les erreurs mesurées ne sont pas les erreurs au sens du sondeur, l'aléa est bien un aléa de **modèle** (ou même plus compliqué avec F&H)

Pour apprécier l'erreur au sens du sondeur, il faut s'en tenir aux approches par simulation (analytique trop compliqué).

Quelques compléments sur l'erreur

1) Penser à l'approche cartographique

On peut juger **visuellement** des corrélations spatiales (naturelles pour Y et v_d).



2) Une méthodologie puissante : l'approche par simulation

Il s'agit de partir d'une population « proche » de notre population d'intérêt et de simuler :

l'échantillonnage,
+
l'estimation « petits domaines »

un grand nombre de fois (1000 ?).

Ensuite, on utilise la loi des grands nombres pour apprécier **le biais ET la MSE** de l'estimateur « petits domaines ».

La population peut être :

- une population réelle disposant d'une variable proche de la variable étudiée (ex : le RP + variable de chômage spontané) ;

- une population totalement artificielle construite via un modèle ;

- un échantillon dupliqué « à la manière du *bootstrap* ».

3) Diagnostic de couverture « avancé »

$$IC(\hat{\theta}_d^{direct}) = \hat{\theta}_d^{direct} \pm k_{0,05} \cdot \sqrt{\hat{V}\hat{\theta}_d^{direct}}$$

$$IC(\hat{\theta}_d^{SAE}) = \hat{\theta}_d^{SAE} \pm k_{0,05} \cdot \sqrt{\hat{V}\hat{\theta}_d^{SAE}}$$

en calculant $k_{0,05} = 2.\left(1 + \frac{\sqrt{\hat{V}\hat{\theta}_d^{direct}}}{\sqrt{\hat{V}\hat{\theta}_d^{SAE}}}\right)^{-1} \cdot \sqrt{1 + \frac{\hat{V}\hat{\theta}_d^{direct}}{\hat{V}\hat{\theta}_d^{SAE}}}$

Si $E\left(\hat{\theta}_d^{SAE}\right) = \theta_d$ alors 95% de ces IC se recoupent.

D'où un test avec la stat. $N_D^{recoupe} \underset{H0}{\rightarrow} B(D;0.95)$

4) Goodness of fit TEST

$$W = \sum_{d=1}^{D} \frac{\left(\hat{\theta}_d^{direct} - \hat{\theta}_d^{SAE}\right)^2}{\underset{p}{\hat{V}}\left(\hat{\theta}_d^{direct}\right) + \underset{\xi}{mse}\left(\hat{\theta}_d^{SAE}\right)}$$

Si $E(\hat{\theta}_d^{SAE}) = \theta_d$ et les n_d « pas trop petits » :

$$W \rightarrow \chi^2(D).$$

En conclusion

- La meilleure stratégie, quand elle est possible, consiste toujours à gonfler en amont la taille de l'échantillon dans le petit domaine !

- La qualité finale reste dépendante de modèles (sauf calage) : pas de miracle ! Comme on a peu d'information locale, il faut compter sur des modèles pour aller en chercher ailleurs…

- Surtout pour les modèles individuels, il peut y avoir un sérieux obstacle dû au manque d'informations auxiliaires explicatives.

- Toute cette mécanique souffre de l'abstraction entourant la notion d'aléa dans un modèle.

- En corollaire, accepter le biais, qui est inévitable.

- L'objectif doit rester humble : *a priori* c'est moins "être bon" que "être meilleur que si on ne fait rien". L'alternative est l'estimation directe, donc souvent la catastrophe, et l'enjeu du choix de méthode est davantage le biais que la variance.

- Le "classement" des méthodes est (très) délicat : modèles différents, aléas différents, multiples critères à prendre en compte (critères d'ajustement de modèle, le biais, la MSE);

- La pertinence des hypothèses (= modèles) continue à s'apprécier essentiellement de manière globale (AIC, graphiques, R^2 si LMM,…) : pour un domaine donné, il n'est pas évident de détecter un modèle mal adapté (surtout pour l'approche non stochastique).

- En terme technique comme en interprétation, la complexité devient beaucoup plus grande avec les variables qualitatives (modèles **GLMM**) :

- On commence à disposer d'outils logiciels spécialisés mais relativement dispersés (pas encore de logiciel *leader* et couvrant large). Les 'package R' se développent : SEA, EMDI – sinon GLMM ou LME4.

- Technique inadaptée à la production de masse : les modèles se construisent variable par variable. Penser néanmoins à l'empilement d'échantillons.

- La technique est extrêmement vaste et n'est pas simple : il faut des moyens en ingénierie statistique.

BIBLIOGRAPHIE

Rao J.N.K., Molina I., *Small Area Estimation*, 2015, WILEY