Modélisation pour l'estimation sur petits domaines

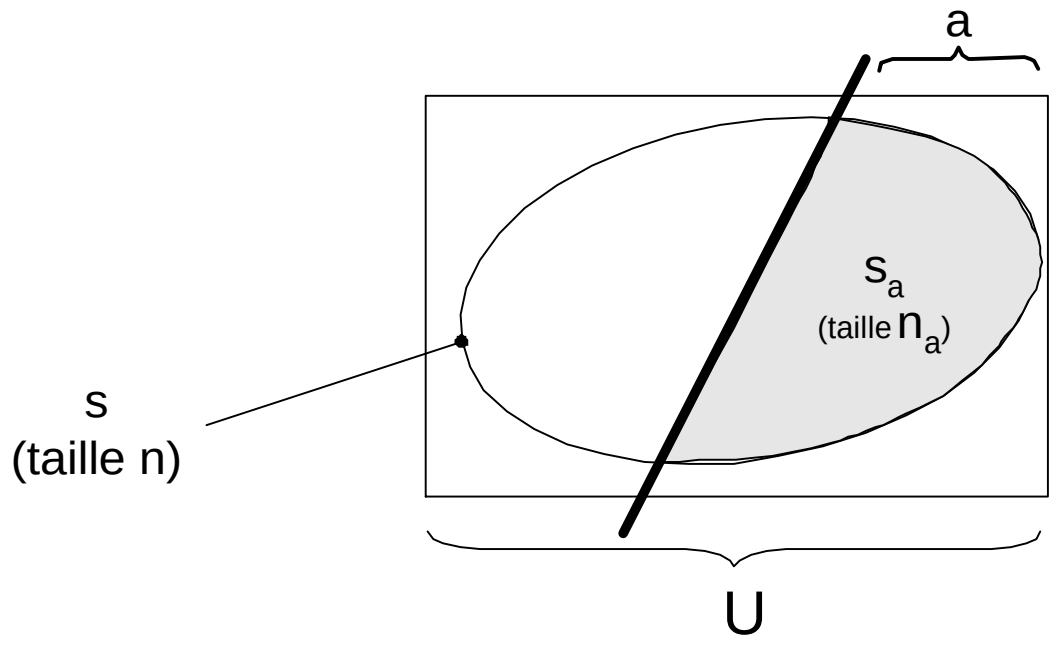
A) *Problématique et définitions*

Population U (taille N)
Domaine = sous population a (taille N_a)

$$Y_a = \sum_{i=1}^{N_a} Y_i \quad \text{et} \quad \bar{Y}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} Y_i \qquad ?$$

Echantillon s tiré dans U (plan complexe)

P_i = probabilité de sélection de l'individu i



Négligeons l'évènement $n_a = 0$

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

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

Donc **pas de problème de biais** !

Mais - hélas :

$$V(\hat{Y}_a) = O\left(\frac{N_a^2}{n_a}\right) \quad \textbf{et} \quad CV\left(\hat{Y}_a\right) = O\left(\frac{1}{\sqrt{n_a}}\right)$$

Donc un sérieux problème d'instabilité si n_a petit !

→ ***3 grandes catégories de méthodes :***

1) Le calage au niveau local

- Utilisation du seul échantillon s_a (n_a « assez grand »)
- Marges LOCALES

⇒ estimation **directe**

2) Modèle impliquant des paramètres NON aléatoires *(approche entièrement descriptive)*

- Modèle = hypothèse simplificatrice de la réalité

  Exemple : $\overline{Y} = \overline{Y}_a$
- Utilisation de l'échantillon COMPLET s
- Pas de calcul (satisfaisant) d'erreur

⇒ estimation **indirecte**

3) Modèle expliquant Y par des variables auxiliaires X (version stochastique : Y aléatoire)

- Utilisation de l'échantillon COMPLET s
- L'effet propre au domaine est isolé et apparaît **explicitement**

⇒ estimation **indirecte**

B) *Pour mémoire : les estimateurs synthétiques (approche descriptive)*

Fondement de l'estimateur synthétique = croire à une **hypothèse descriptive** du type

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

Généralement

$$\hat{Y}_{Reg,a} = \hat{Y}_a + \hat{B}_a^T \left(X_a - \hat{X}_a \right) = X_a^T \hat{B}_a$$

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

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

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

Vision stochastique : $Y_i = B^T . X_i + e_i$ avec $E(e_i) = 0$.

Les spécificités locales sont prises en compte, mais « seulement » par l'intermédiaire des variables auxiliaires X_i.

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

$\tilde{B} =$ vrai coefficient de régression (population finie)

Pas possible d'estimer de manière <u>stable</u> les EQM des estimateurs synthétiques (c'est **la composante de biais qui pose un problème de fond !)**.

Elargissement à la classe des estimateurs composites :

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

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

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

C) *Typologie des estimateurs indirects avec modélisation explicite*

On distingue :

- L'unité statistique modélisée :
  - Modèle au niveau du **domaine**
  - Modèle au niveau des **individus**.

- Différentes familles d'estimateurs, selon le contexte :
  - Les estimateurs sans biais linéaires optimaux
  - Les estimateurs de comptage
  - Les estimateurs optimaux
  - Les estimateurs Bayésiens.

On s'intéresse toujours à **m** domaines : a varie de 1 à m : c'est seulement dans ce contexte qu'on « tirera de la force » d'une modélisation (« *to borrow strenght* »)

Les modèles stochastiques traitant les variables **quantitatives** sont de type « **modèle linéaire général** » :

$$\boxed{Y = X \cdot \beta + \varepsilon}$$

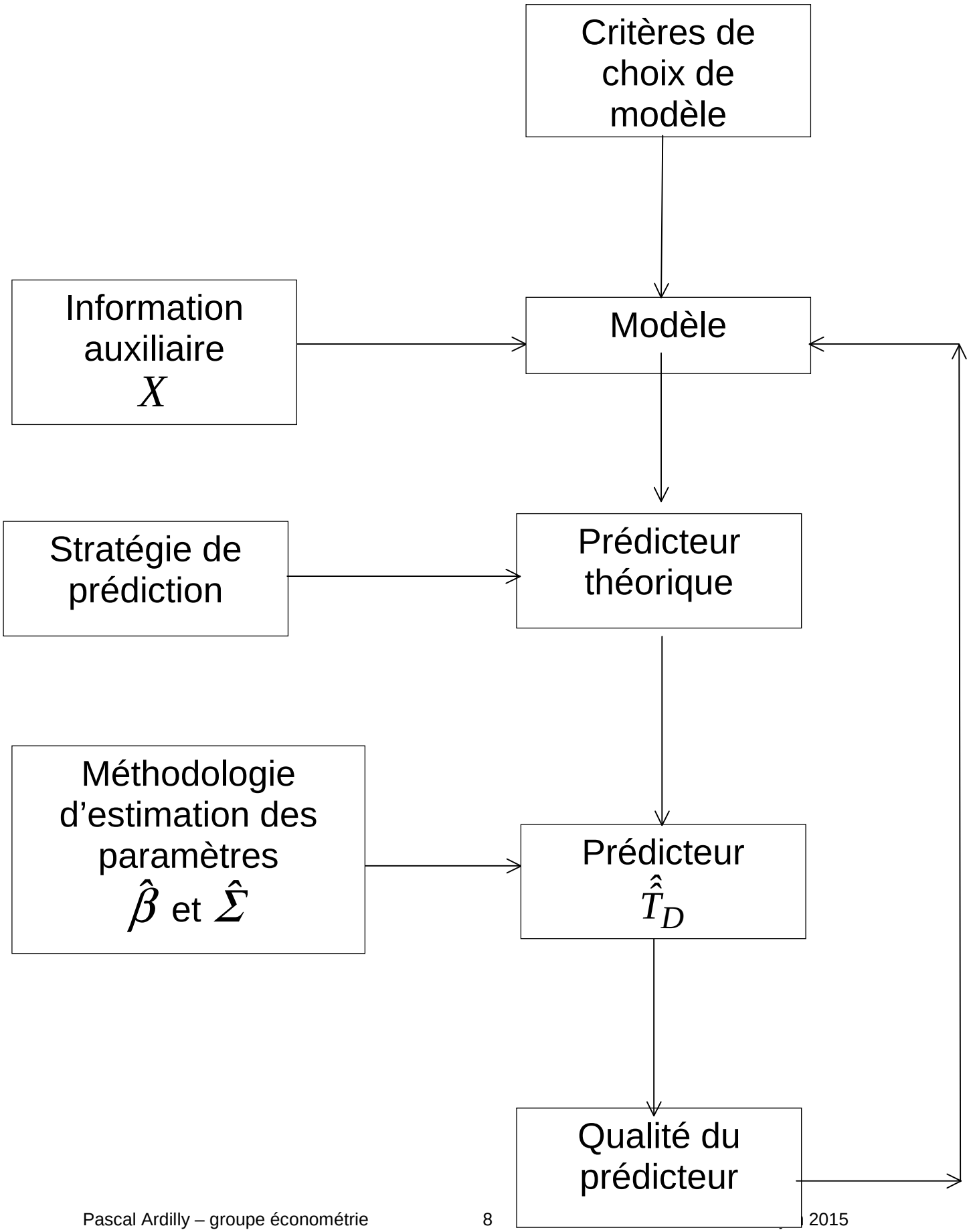
$$E\varepsilon = 0 \qquad \text{et} \qquad Var\varepsilon = \Sigma$$

Idée centrale :

Ce modèle est valable de la même façon sur toute la population. Il va servir à prédire la variable aléatoire $Y_a = \sum_{i=1}^{N_a} Y_i$ **au moyen d'une estimation (très) fiable des paramètres** β **et** Σ **(fiable en ce sens où elle s'appuie sur l'échantillon complet** s**).**

Les modèles stochastiques traitant les variables **qualitatives** sont des **modèles linéaires généralisés** (ou mixtes généralisés).

D) *Choix de méthode et modèle*



E) Estimateurs BLUP dans le modèle individuel (Battese, Harter et Fuller)

La variable d'intérêt Y est **quantitative** **et** **continue**.

$$Y_{a,i} = X^{t}_{a,i} \cdot \beta + \nu_a + e_{a,i}$$

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

$X_{a,i}$ = vecteur d'effets fixes - quantitatifs / qualitatifs.

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

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

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

En toute généralité, pas de lois pour les aléas (mais ça aide bien quand même !)

C'est un modèle linéaire mixte

→ *One-fold nested error linear regression model*
→ *Modèle multi niveaux*

Nota : pas de différence de nature entre ν et e sur le plan *technique* – mais des différences d'interprétation !

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

$$Y = X\beta + v + e = X\beta + \varepsilon \qquad \text{avec } Var(\varepsilon) = V$$

$$V = \begin{pmatrix} \Sigma_1 & 0 & 0 & 0 \\ 0 & \Sigma_2 & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & \Sigma_m \end{pmatrix} \quad \text{où}$$

$$\Sigma_k = \begin{pmatrix} \sigma_v^2 + \sigma_e^2 & \sigma_v^2 & \sigma_v^2 \\ \sigma_v^2 & ... & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 & \sigma_v^2 + \sigma_e^2 \end{pmatrix} \qquad \forall k = 1,2,...,m$$

Le paramètre à estimer $\overline{Y}_a$ (vision 'sondeur') est une **variable aléatoire à prédire.** C'est donc une approche originale pour le sondeur.

Il faut que le modèle - valide sur la population complète - reste valide à l'identique sur S

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

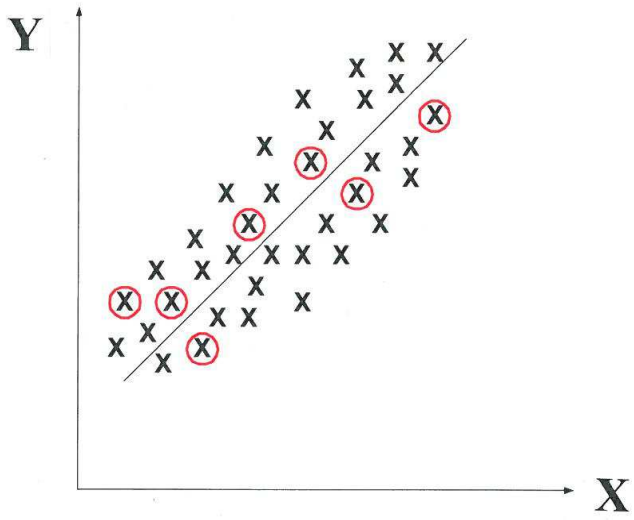
$\tilde{\vec{X}}_i$ = variables utilisées pour tirer S (ex : critère de taille, stratification, équilibrage,...)

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

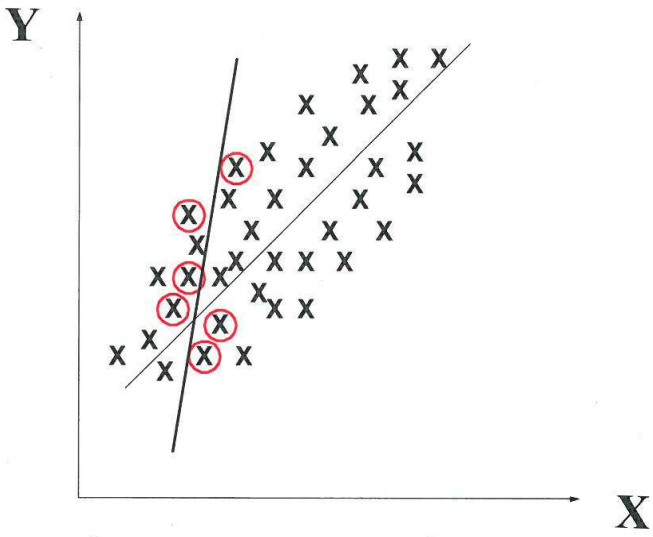
Par la suite, on **prétend** qu'il existe un mécanisme universel qui s'applique au-delà de la population (super-population) et

- les modèles sont supposés bien spécifiés
- l'échantillonnage est supposé non informatif

En rouge, échantillon **s** dans **U** (ou $\mathbf{s}_a$ dans $\mathbf{U} \cap \mathbf{a}$)



Contexte 1 : cas favorable



Contexte 2 : cas défavorable

Propriété **toujours vraie avec un sondage aléatoire simple.**

Une solution **:** inclure $\tilde{X}_a$ dans le vecteur X_a.

→ ***Application de la stratégie de prédiction***

On cherche le prédicteur $\hat{\bar{Y}}_a^H$ vérifiant :

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

$$E\left(\hat{\bar{Y}}_a^H - \bar{Y}_a\right) = 0$$

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

La solution est dite BLUP

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

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

Supposons $n_a << N_a$

L'estimateur solution BLUP de $\bar{Y}_a$ 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 $\tilde{Y}_{a,i} = X_{a,i} \cdot \tilde{\beta} + \tilde{\nu}_a$ prédicteur individuel optimum

avec :

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

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

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

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

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

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

Pratiquement, l'information individuelle $X_{a,i}$ provient du questionnaire et son vrai total d'une source exhaustive (base de sondage ou pas).

ATTENTION : risque d'hétérogénéité !!!

DONC : l'information auxiliaire doit être :

a) explicative de la variable d'intérêt ;

b) présente dans le questionnaire ;

c) présente dans la source auxiliaire - et relever exactement du même concept.

On a également

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

Partie (pseudo) directe Partie synthétique

Cet estimateur est dit « **composite** ».

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

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

n_a grand $\Rightarrow$ partie directe prime

Avec seulement l'aléa de sondage, cet estimateur n'a pas de bonnes propriétés (biais, non convergent) : mais **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 $\rightarrow$ estimateur empirique EBLUP (E=empirique) - qui est le seul "véritable estimateur / prédicteur".

On peut utiliser :

- La méthode des moments
- L'estimation du maximum de vraisemblance (EMV), (postuler une **loi de Gauss** pour e et v) : pas de solution analytique, mais une solution numérique.

Pourquoi pas un modèle à effets fixes ?

$$Y_{a,i} = X_{a,i}^t \cdot \beta + v_a + e_{a,i}$$

avec v_a **de même nature** que β.

C'est possible !

C'est même le plus naturel quand a est un domaine. :

- qui n'est pas une sous-population de constitution « aléatoire » ;
- pour laquelle v_a est 'aussi intéressant' que β.

L'hypothèse devient : $EY_{a,i} = X_{a,i}^t \cdot \beta + v_a$

Techniquement, aucun problème : modèle **d'analyse de la covariance** classique.

Le (seul mais gros) problème c'est que

$$V(\hat{v}_a) = O\left(\frac{1}{n_a}\right)$$

et du coup $V(\hat{\bar{Y}}_a) = O\left(\frac{1}{n_a}\right)$: problème initial !!!

→ Erreur et estimation de l'erreur de **l'EBLUP**

$$EQM\left(\hat{\bar{Y}}_a^{EBLUP}\right)=g_1(\sigma_v^2,\sigma_e^2)+g_2(\sigma_v^2,\sigma_e^2)+g_3(\sigma_v^2,\sigma_e^2)$$

$$\hat{eqm}\left(\hat{\bar{Y}}_a^{EBLUP}\right)=g_1(\hat{\sigma}_v^2,\hat{\sigma}_e^2)+g_2(\hat{\sigma}_v^2,\hat{\sigma}_e^2)+2.g_3(\hat{\sigma}_v^2,\hat{\sigma}_e^2)$$

avec

$$g_1(\sigma_v^2,\sigma_e^2)=\gamma_a\cdot\frac{\sigma_e^2}{n_a}$$

$$g_2(\sigma_v^2,\sigma_e^2)=\left(\bar{X}_a-\gamma_a\cdot\bar{x}_a\right)^t\cdot V\tilde{\beta}\cdot\left(\bar{X}_a-\gamma_a\cdot\bar{x}_a\right)$$

$$g_3(\sigma_v^2,\sigma_e^2)=\frac{1}{n_a^2}\cdot\frac{1}{\left(\sigma_v^2+\frac{\sigma_e^2}{n_a}\right)^3}$$
$$\cdot\left(\sigma_e^4\cdot V_\infty(\hat{\sigma}_v^2)+\sigma_v^2\cdot V_\infty(\hat{\sigma}_e^2)-2\cdot\sigma_v^2\cdot\sigma_e^2\cdot Cov_\infty(\hat{\sigma}_v^2,\hat{\sigma}_e^2)\right)$$

Conclusion :

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

F) Une variante : les estimateurs pseudo-EBLUP

L'EBLUP n'est (même) pas « *design-consistent* ». Pour y remédier, on insère les POIDS de sondage $w_{a,i}$ dans l'estimateur :

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

$$\frac{\sum_{i \in s_a} w_{a,i} \cdot Y_{a,i}}{\sum_{i \in s_a} w_{a,i}} = \frac{\sum_{i \in s_a} w_{a,i} \cdot X_{a,i}}{\sum_{i \in s_a} w_{a,i}} \cdot \beta + v_a + \frac{\sum_{i \in s_a} w_{a,i} \cdot e_{a,i}}{\sum_{i \in s_a} w_{a,i}}$$

Soit $\bar{y}_{a,w} = \bar{x}_{a,w}^T \cdot \beta + v_a + \bar{e}_{a,w}$

avec $$V(\bar{e}_{a,w}) = \sigma_e^2 \cdot \frac{\sum_{i \in s_a} w_{a,i}^2}{\left(\sum_{i \in s_a} w_{a,i} \right)^2} = \sigma_e^2 \cdot \delta_{aw}^2$$

$$\boxed{\hat{\bar{Y}}_a^{ps\_BLUP} = \bar{X}_a^t \cdot \tilde{\beta}_w + \gamma_{a,w}\left(\bar{y}_{a,w} - \bar{x}_{a,w\cdot}^t \cdot \tilde{\beta}_w\right)}$$

$$\tilde{\beta}_w = \left(\sum_{a=1}^{m} \left(\sum_{i\in s_a} w_{a,i} \cdot X_{a,i} \cdot X_{a,i}^T - \gamma_{a,w} \cdot \left(\sum_{i\in s_a} w_{a,i}\right) \cdot \bar{x}_{a,w} \cdot \bar{x}_{a,w}^T\right)\right)^{-1} \times$$

$$\left(\sum_{a=1}^{m} \left(\sum_{i\in s_a} w_{a,i} \cdot X_{a,i} \cdot Y_{a,i} - \gamma_{a,w} \cdot \left(\sum_{i\in s_a} w_{a,i}\right) \cdot \bar{x}_{a,w} \cdot \bar{y}_{a,w}\right)\right)$$

$$\gamma_{a,w} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2 \cdot \delta_{aw}^2}$$

$$\boxed{\hat{\bar{Y}}_a^{ps\_EBLUP} = \bar{X}_a^t \cdot \hat{\beta}_w + \hat{\gamma}_{a,w}\left(\bar{y}_{a,w} - \bar{x}_{a,w\cdot}^t \cdot \hat{\beta}_w\right)}$$

Propriétés :

i) Si n_a « grand », $\delta_{aw}^2 = O\left(\frac{1}{n_a}\right) \to 0$ et $\gamma_{a,w} \to 1$, donc

$\bar{x}_{a,w} \approx \bar{X}_a$ et $\hat{\bar{Y}}^{ps\_EBLUP} \approx \bar{y}_{a,w} \approx \bar{Y}_a$: *design-consistent*

ii) Si $\forall a = 1,2,...m$: $\sum_{i\in s_a} w_{a,i} = N_a$ (calage), alors

$$\sum_{a=1}^{m} N_a \cdot \hat{\bar{Y}}_a^{ps\_EBLUP} = \hat{Y}_{HT} + \hat{\beta}_w^t \cdot \left(X - \hat{X}_{HT}\right)$$

Noter une variante du Pseudo-EBLUP : ne pas utiliser les poids dans l'estimateur de β :

$$\hat{\beta}_w = \left(\sum_{a=1}^{m} \left(\sum_{i \in s_a} w_{a,i} \cdot X_{a,i} \cdot X_{a,i}^T - \hat{\gamma}_{a,w} \cdot \left(\sum_{i \in s_a} w_{a,i} \right) \cdot \bar{x}_{a,w} \cdot \bar{x}_{a,w}^T \right) \right)^{-1} \times$$

$$\left(\sum_{a=1}^{m} \left(\sum_{i \in s_a} w_{a,i} \cdot X_{a,i} \cdot Y_{a,i} - \hat{\gamma}_{a,w} \cdot \left(\sum_{i \in s_a} w_{a,i} \right) \cdot \bar{x}_{a,w} \cdot \bar{y}_{a,w} \right) \right)$$

(re)devient

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

$$\boxed{\hat{\bar{Y}}_a^{ps\_EBLUP} = \bar{X}_a^t \cdot \hat{\beta} + \hat{\gamma}_{a,w} \left(\bar{y}_{a,w} - \bar{x}_{a,w\cdot}^t \cdot \hat{\beta} \right)}$$

C'est un scénario « batard ».

Les 2 pseudo EBLUP restent biaisés (aléa de sondage), comme l'EBLUP. Leurs EQM ont une allure proche de celle de l'EBLUP.

Les EQM des pseudo EBLUP **ne prennent pas en compte la variabilité due à l'échantillonnage** : il s'agit d'erreurs dues au modèle, **conditionnellement à l'échantillon**.

Diagnostics de modèle

On cherche à valider le modèle, à partir de tests et de graphiques.

***Ca n'assure pas que le modèle est valide sur un domaine donné**, il n'y a jamais de preuve d'absence d'un comportement local spécifique dès que* n_a *est petit*.

On peut ainsi voir si l'hypothèse est acceptable au **niveau global**.

Les tests classiques ne semblent pas éclairants pour détecter des spécificités de comportements locaux :

$$\begin{pmatrix} \vec{y}_1 \\ \vec{y}_2 \\ \\ \vec{y}_D \end{pmatrix} \rightarrow N\left(\begin{pmatrix} X_1 & & & \\ & X_2 & & \\ & & \dots & \\ & & & X_D \end{pmatrix} . \begin{pmatrix} B_1 \\ B_2 \\ \dots \\ B_D \end{pmatrix}, \Sigma \right)$$

$$H_0 : \forall a \quad B_a = B \quad ???$$

Mais $Var\left(\hat{B}_a\right) = O\left(\frac{1}{n_a}\right)$

$\Rightarrow$ on va très (trop) souvent accepter H_0.

$\rightarrow$ on peut quand même tester : si on rejette H_0 c'est vraiment qu'il y a un effet local !

$\rightarrow$ il faudrait néanmoins une source externe pour être au clair

F) Une modélisation au niveau domaine prenant en compte le plan de sondage : le modèle de Fay & Herriot

$$\bar{Y}_a = X_a^t \cdot \beta + b_a \cdot v_a$$

Ref : Fay, Herriot (JASA, 1979)

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

v_a (« 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 encore que $\bar{Y}_a$ *est* ***quantitative*** *et de* ***nature continue***. **On applique néanmoins ce modèle pour des proportions** P_a **ou des dénombrements si** N_a **est 'grand'**.

$$\hat{\bar{Y}}_a = \bar{Y}_a + e_a$$

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

$$\hat{\bar{Y}}_a = X^t_a \beta + b_a v_a + e_a$$

On considérera v_a et e_a comme indépendantes.

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

$\sigma_v^2 \rightarrow$ **variabilité de type « inter » domaines**

$\psi_a \rightarrow$ **variabilité « intra » domaines**.

La modélisation considérant v_a comme effet FIXE est plus naturelle mais le modèle n'est plus estimable.

C'est un modèle linéaire mixte,

qui mêle 2 natures d'aléas

$\bar{Y}_a$ est une **variable aléatoire** à prédire.

La stratégie BLUP conduit à :

$$\boxed{\hat{\bar{Y}}_a^H = X_a^t \tilde{\beta} + \gamma_a \left(\hat{\bar{Y}}_a - X_a^T \tilde{\beta} \right)}$$

ou encore

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

avec

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

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

De nouveau un estimateur **composite** !

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

$$V(\tilde{\beta})=O\left(\frac{1}{\sum_{a=1}^{m} n_a}\right)$$

* 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 en général que $\hat{\bar{Y}}_a^H = X_a^T \tilde{\beta}$.

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

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

2) On peut imposer des b_a différenciés pour permettre à la variance de $\overline{Y}_a$ de dépendre de a :

$$Var\overline{Y}_a = b_a^2 \cdot \sigma_v^2$$

(pas 'spontanément'… mais une fois détectée la présence d' « outliers »)

RAPPEL : le biais n'est égal à 0 que si on considère les 2 aléas conjointement (modèle + échantillonnage).

Il n'est pas habituel d'utiliser un modèle où la variable expliquée $\hat{\overline{Y}}_a$ peut être très différente de la réalité $\overline{Y}_a$!

Noter qu'il n'y a pas de question d'homogénéité d'information auxiliaire comme dans le modèle de *Battese / Fuller* - ici la « bonne » corrélation entre X_a et Y_a suffit.

En pratique, on a souvent une (bien) plus grande richesse d'information au niveau agrégé qu'au niveau individuel.

Comme σ_v^2 est inconnu, il faut l'estimer pour construire l'estimateur EBLUP / ESBOL.

Il y a plusieurs méthodes d'estimation de σ_v^2 :
la plus courante = EMV (ou REMV)

Conclusion

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 :

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

La troncature va réduire de fait la variance.

→ ERREUR :

$$EQM\left(\hat{\bar{Y}}_a^{BLUP}\right) = \gamma_a \cdot \psi_a + (1-\gamma_a)^2 \cdot X_a^t \cdot \left(\sum_{a=1}^{m} \frac{X_a \cdot X_a^T}{\psi_a + b_a^2 \cdot \sigma_v^2}\right)^{-1} \cdot X_a$$

$$EQM\left(\hat{\bar{Y}}_a^{BLUP}\right) = g_1(\sigma_v^2) + g_2(\sigma_v^2)$$

Si m **est grand :** $$\frac{EQM(\hat{\bar{Y}}_a^{BLUP})}{EQM(\hat{\bar{Y}}_a)} \approx \gamma_a$$

Si γ_a **petit** ⇒ **gain important**.

$$EQM\left(\hat{\bar{Y}}_a^{EBLUP}\right) = E\left(\hat{\bar{Y}}_a^{EBLUP} - \bar{Y}_a\right)^2$$

$$EQM\left(\hat{\bar{Y}}_a^{EBLUP}\right) = g_1(\sigma_v^2) + g_2(\sigma_v^2) + g_3\left(\sigma_v^2\right)$$

$$g_3(\sigma_v^2) \approx \frac{b_a^4 \cdot \psi_a^4}{\left(\psi_a + b_a \cdot \sigma_v^2\right)^3} \cdot Var_\infty\left(\hat{\sigma}_v^2\right) = O\left(\frac{1}{m}\right)$$

Estimation de l'erreur de l'EBLUP

$$\hat{mse}\left(\hat{\bar{Y}}_a^{EBLUP}\right) \approx g_0(\hat{\sigma}_v^2) + g_1(\hat{\sigma}_v^2) + g_2(\hat{\sigma}_v^2) + 2 \cdot g_3\left(\hat{\sigma}_v^2\right)$$

$g_0(\sigma_v^2)$: expression complexe en $\frac{1}{m}$ (fonction de la méthode d'estimation de σ_v^2).

Rappel : si m petit, on ne saura pas estimer l'erreur !

G) Deux familles de modèles plus complexes (par exemple)

- ***Modèles de corrélation spatiale***

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

Autre approche : Ω_a étant un « voisinage » de a,

$$v_a \left| \{v_b, b \neq a\} \rightarrow \mathcal{N}\left(\rho \cdot \sum_{b \in \Omega_a} v_b \, , \sigma_v^2 \right) \right.$$

- ***Modèles temporels***

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

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

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

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).

H) Les estimateurs de comptage (variables qualitatives)

a) Modélisation au niveau de l'individu : exemple du modèle Logistique avec effet aléatoire

Modèle :

$$Y_{a,i} \rightarrow B\left(1, P_{a,i}\right)$$

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

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

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

Les $Y_{a,i}$ sont deux à deux indépendantes **conditionnellement à** ν_a. Mais la densité jointe des $Y_{a,i}$ est très complexe : **l'estimation par EMV n'est plus possible.**

On transforme le modèle initial (par linéarisation) en un **modèle linéaire mixte « approché ».**

$$E(Y_{a,i} \mid \nu_a) = P_{a,i} = g^{-1}(\eta) \quad \text{où} \quad \eta = X_{a,i}^{T} \cdot \beta + \nu_a$$

Soit une valeur "fixe" $\tilde{\eta} = X_{a,i}^{T} \cdot \tilde{\beta} + \tilde{v}_a$ et la nouvelle variable observée :

$$L_{a,i} = \tilde{\Delta}^{-1}\left(Y_{a,i} - g^{-1}(\tilde{\eta})\right) + X\tilde{\beta} + \tilde{v}_a \ ,$$

$$\tilde{\Delta} = \left.\frac{\partial g^{-1}}{\partial \eta}\right|_{\eta=\hat{\eta}}$$

$$E(L_{a,i} | v_a) = X_{a,i}^{T} \cdot \beta + v_a$$

avec $V(L_{a,i} | v_a) = \tilde{\Delta}^{-1} V(Y_{a,i} | v_a) \tilde{\Delta}^{-1}$

Le modèle linéaire mixte approché est donc :

$$L_{a,i} = X_{a,i}^{T} \cdot \beta + v_a + \varepsilon_{a,i}$$

où $V\varepsilon_{a,i} = \sigma_{\varepsilon}^{2} \cdot \tilde{\Delta}^{-1} \cdot \tilde{\Delta}^{-1}$ (*à voir...*)

σ_v^2 est estimé par maximum de vraisemblance.

D'où les composantes EBLUP $\hat{\beta}$ et $\hat{v}_a$.

$$\Rightarrow \quad \hat{N}_a^{c,H} = \sum_{\substack{i \in s \\ i \in a}} 1_{i \in c} + \sum_{\substack{i \notin s \\ i \in a}} g^{-1}(X_{a,i}^T \hat{\beta} + \hat{v}_a)$$

On obtient donc des estimateurs « d'inspiration » BLUP, donc de nature synthétique et cela va limiter leur variance.

b) Le modèle de Poisson avec effet aléatoire

Paramètre : N_a^c (ou proportion associée).

Modèle : $\hat{N}_a^c$ (estimateur direct) $\longrightarrow \mathrm{P}(\lambda_a)$ avec

$$F(\lambda_a) = z_a^T \beta + \nu_a$$

La fonction *F* est souvent un logarithme.

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

On passe par un modèle linéaire mixte approché, on obtient $\hat{\beta}$ et $\hat{\nu}_a$, puis in fine

$$\hat{N}_a^{c,H} = F^{-1}(z_a^T \hat{\beta} + \hat{\nu}_a)$$

Possibilité de modéliser le cas $Var\ \hat{N}_a^c > E\ \hat{N}_a^c$ ("overdispersion" : $\hat{N}_a^c \xrightarrow[Loi]{}$ famille exponentielle)

I) Les estimateurs optimaux

A) Principe général et démarche d'ensemble :

On veut :

- que l'optimum soit « absolu » (pas seulement parmi la classe des estimateurs linéaires) ;
- que la théorie s'applique aux variables quantitatives comme qualitatives ;

Il faut une **hypothèse sur les lois de e et de v.**

* Théorème central (rappel):

Pour prédire une variable aléatoire μ au moyen d'une variable aléatoire Y, le prédicteur optimum au sens de l'erreur quadratique est

$$f(Y) = E[\mu|Y]$$

Alors pour tout g, on a $E[g(Y)-\mu]^2 \geq E[f(Y)-\mu]^2$

Théorème :

Si $Y = X\beta + Zv + e$, si $\mu = l^T \cdot \beta + m^T \cdot v$
et si $(v, e) \rightarrow$ Gauss, alors

$$f(Y) = l^T \beta + m^T V(v) \cdot Z^T V^{-1}(Y - X\beta)$$

en posant $V = Z \cdot V(v) \cdot Z^T + V(e)$

Démarche à suivre

a/ Soit

- La densité de μ : $f(\mu\,;\lambda_2)$
- La densité de Y sachant μ : $f(Y|\mu\,;\lambda_1)$

b/ Par la formule de Bayes :

$$f(\mu|Y\,;\lambda_1,\lambda_2)=\frac{f(Y|\mu\,;\lambda_1)\cdot f(\mu\,;\lambda_2)}{\int f(Y|\mu;\lambda_1)\cdot f(\mu\,;\lambda_2)d\mu}$$

c/ Puis

$$\hat{\mu}^{OPTI}=E(\mu|Y\,;\lambda_1,\lambda_2)=\int \mu\cdot f(\mu|Y\,;\lambda_1,\lambda_2)d\mu$$

C'est le prédicteur optimum théorique (incalculable en général car on ne connaît pas λ_1 ni λ_2).

d/ On estime (λ_1,λ_2) à partir de

$$f(Y\,;\lambda_1,\lambda_2)=\int f(Y|\mu\,;\lambda_1)\cdot f(\mu\,;\lambda_2)d\mu$$

par une méthode quelconque (par exemple le maximum de vraisemblance).

e/ On obtient $(\hat{\lambda}_1,\hat{\lambda}_2)$ et on termine en calculant

$$E(\mu|Y,\hat{\lambda}_1,\hat{\lambda}_2)\;=\;\hat{\mu}_E^{OPTI}$$

$\hat{\mu}_E^{OPTI}$ est dit abusivement estimateur « Bayésien empirique ». En fait, il n'y a rien de « Bayésien » dans cette affaire !

B) Application à un paramètre de « proportion »

Objectif : estimer des vraies **proportions** P_a

Soit : $Y_{a,i}$ = 1 si $(a,i) \in D$ (une sous-pop.)
= 0 sinon

Supposons : $Y_{a,i} \to \mathcal{B}(1, P_a)$

On suppose aussi les $Y_{a,i}$ indépendants d'un individu à l'autre dans un domaine donné.

$$Y_a = \sum_{i \in s_a} Y_{a,i} \qquad \Rightarrow \qquad Y_a \to \mathcal{B}(n_a, P_a)$$

$\mu = P_a$ (à prédire) et $Y = Y_a$ (observé)

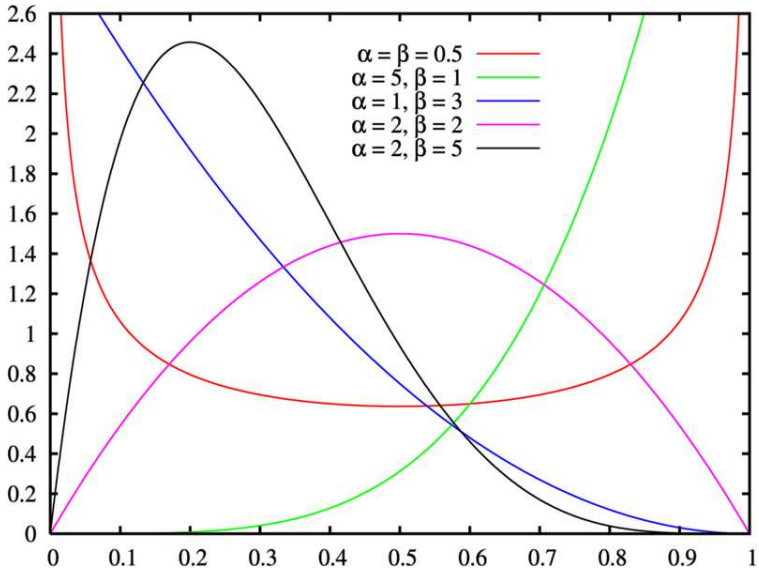
Nota : pas de poids de sondage : peu importe la méthode de tirage de s_a.

A partir de là, on peut imaginer

nombre de modèles sur P_a !!!

- Exemple 1 : P_a suit une **loi bêta** (α, β)

$$f(P_a\,;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}P_a^{\alpha-1}(1-P_a)^{\beta-1}$$



$$f(Y_a|P_a)=\binom{n_a}{Y_a}P_a^{Y_a}\cdot(1-P_a)^{n-Y_a}$$

on déduit (étape b /)

$$f(P_a|Y_a;\alpha,\beta)=\frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+Y_a)\Gamma(n_a-Y_a+\beta)}P_a^{\alpha+Y_a-1}(1-P_a)^{n_a-Y_a+\beta-1}$$

= loi béta $(\alpha+Y_a, n_a-Y_a+\beta)$

Prédicteur optimum (étape c /) :

$$\hat{P}_a^{OPTI} = E\left(P_a \middle| Y_a\ ; \alpha, \beta\right) = \frac{Y_a + \alpha}{n_a + \alpha + \beta}$$

A rapprocher de $\hat{p}_a = \frac{Y_a}{n_a}$ (cas du SAS)

L'étape d/ fournit la loi marginale de Y_a (bêta-binomiale) :

$$f\left(Y_a\ ; \alpha, \beta\right) = \binom{n_a}{Y_a} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+Y_a)\cdot\Gamma(\beta+n_a-Y_a)}{\Gamma(\alpha+\beta+n_a)}$$

D'où $\hat{\alpha}$ et $\hat{\beta}$, estimateurs du maximum de vraisemblance.

Pas de solutions analytiques $\Rightarrow$ utilisation d'algorithmes convergents. On aboutit au prédicteur optimum empirique :

$$\hat{P}_{a,E}^{OPTI} = \frac{Y_a + \hat{\alpha}}{n_a + \hat{\alpha} + \hat{\beta}}$$

Alternative : estimateurs de α et β selon la méthode des moments. Dans ce cas :

$$\hat{P}_{a,E}^{OPTI} = \hat{\gamma}_a \cdot \hat{p}_a + (1-\hat{\gamma}_a) \cdot \hat{p}$$

Direct Synthétique

où $$\hat{p} = \sum_{a=1}^{m} \frac{n_a}{n} \cdot \hat{p}_a = \frac{\sum_{a=1}^{m} Y_a}{n}$$

$$\hat{\gamma}_a = \frac{n_a}{n_a + \hat{\alpha} + \hat{\beta}} \quad \in [0,1].$$

Nota : la variance d'échantillonnage n'intervient jamais dans $\hat{\gamma}_a$, ce qui est un atout très appréciable (et qui n'avait pas lieu avec les modèles linéaires mixtes).

- Exemple 2 :

$$Log \frac{P_a}{1-P_a} = \mu + v_a$$

où $v_a \rightarrow \mathcal{N}(0, \sigma^2)$

PAS d'expression analytique de $\hat{P}_a^{OPTI}$!!!

On peut néanmoins traiter la question par une méthode approchée s'appuyant sur des simulations et sur la loi des grands nombres.

→ *Retour au cas du modèle de Fay et Herrict*

On reprend le modèle de Fay et Herriot en introduisant une hypothèse de normalité, soit :

$$\hat{\theta}_a = \theta_a + e_a \qquad \text{avec } e_a \to \mathcal{N}(0, \Psi_a)$$
$$\theta_a = z_a^T \beta + b_a v_a \quad \text{avec } v_a \to \mathcal{N}(0, \sigma_v^2)$$

On vérifie

$$f\left(\theta_a \middle| \hat{\theta}_a ; \beta, \sigma_v^2\right) \to \mathcal{N}\left(\hat{\theta}_a^{OPTI}, \gamma_a \Psi_a\right)$$

où $$\gamma_a = \frac{b_a^2 \sigma_v^2}{b_a^2 \sigma_v^2 + \Psi_a}$$

Ψ_a est supposé connu, et

$$\hat{\theta}_a^{OPTI} = E\left[\theta_a \middle| \hat{\theta}_a ; \beta, \sigma_v^2\right] = \gamma_a \hat{\theta}_a + (1-\gamma_a) z_a^T \beta .$$

Pour estimer β et σ_v^2 on utilise :

$$\hat{\theta}_a \to \mathcal{N}\left(z_a^T \beta, b_a^2 \sigma_v^2 + \Psi_a\right)$$

Par EMV, on obtient $\hat{\beta}$ et $\hat{\sigma}_v^2$. Finalement :

$$\boxed{\hat{\theta}_{a,E}^{OPTI} = \hat{\gamma}_a \hat{\theta}_a + (1-\hat{\gamma}_a) z_a^T \hat{\beta}}$$

C'est le même estimateur que l'EBLUP (donc il y a une robustesse à l'hypothèse de normalité) !

I) Mémento sur la qualité des estimations

Comment évaluer la qualité des estimations « petits domaines » ??? On ne peut avoir que des présomptions : **il y a toujours « quelque part » un acte de foi**.

a) **Estimer l'EQM** (c'est plus fidèle avec la modélisation explicite) ;

b) Utiliser les **diagnostics de qualité des modèles** en cas de modélisation explicite :

- sélection de variables explicatives ;
- graphique des résidus (utilisation éventuelle de cartes si domaines de nature géographique) ;
- qualité de l'ajustement (ex : AIC, BIC,…) ;
- détection des individus influents.

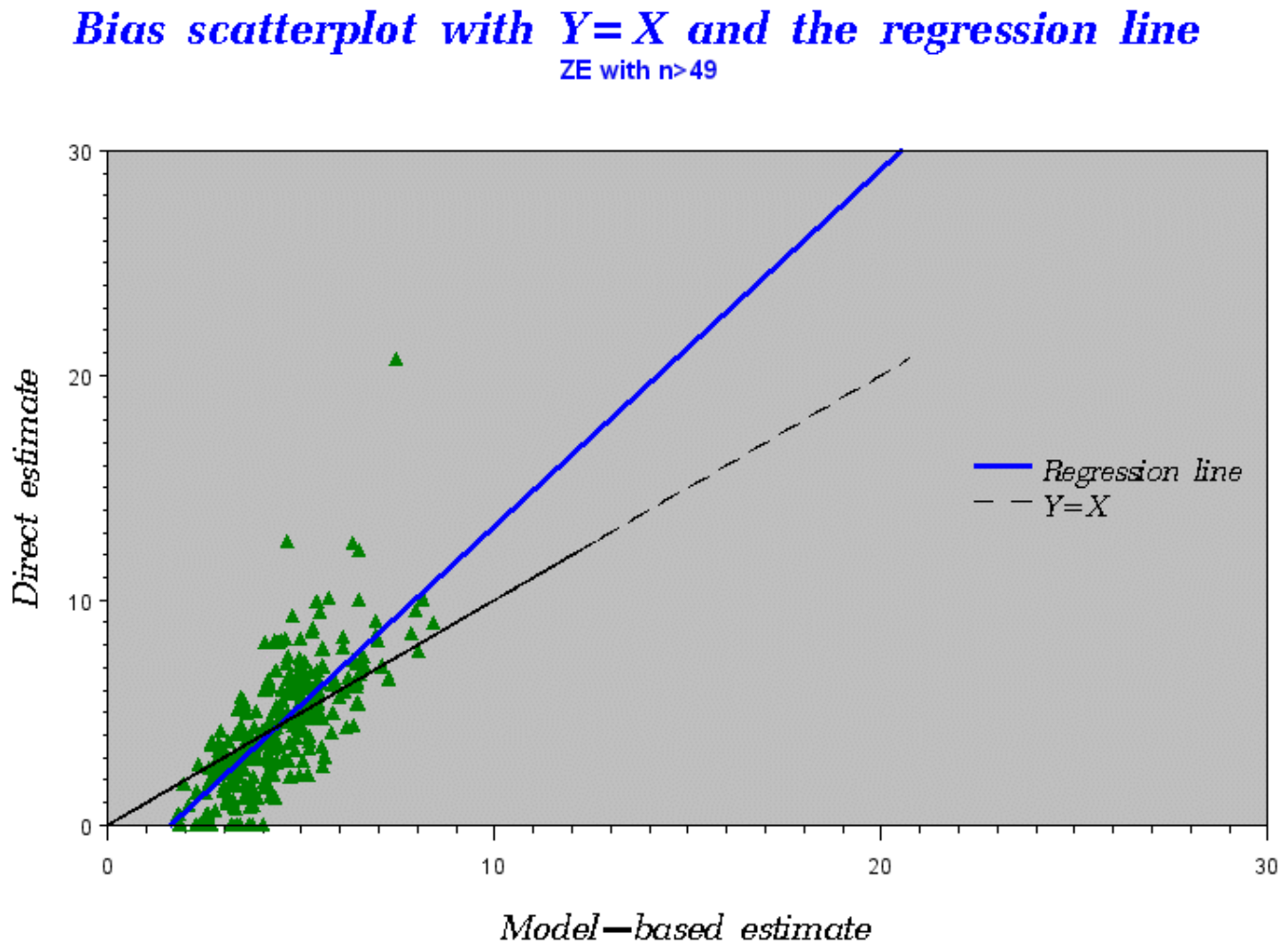
c) Lorsque c'est possible, **comparer à des estimateurs directs** ressentis comme *a priori* fiables (cas d'extension locale par exemple).

Il y a néanmoins un artefact : en cas d'extension, l'estimateur mixte est *par construction* proche de l'estimateur direct ! Il n'y a donc aucune « garantie ». L'estimateur synthétique ne souffre pas de ce phénomène pervers.

d) **Sommer les estimations « petits domaines » sur un « grand » domaine**, puis apprécier l'écart à l'estimateur direct (réputé fiable).

e) **Régresser les estimateurs directs sur les estimateurs « petits domaines »** :

répartition **<u>symétrique</u>** attendue autour de $y = x$ si les estimateurs « petits domaines » ont un faible biais. Seule compte l'allure « générale » du graphique – et non la dispersion du nuage des points.



Shrinkage :

phénomène d'uniformisation consécutif à une modélisation : une partie de la variabilité n'est pas reproduite par l'estimation de type synthétique.

Conséquence : « écrasement » de la distribution des $\hat{Y}_d^{SAE}$ par rapport à celle des $\hat{Y}_d$;

Test de shrinkage :

H_0 : pente de régression = 1

symétrie $\Rightarrow$ pas de biais

réciproque douteuse

f) Par sécurité : **mettre en œuvre plusieurs méthodes d'estimation « petits domaines »** et comparer les distributions des estimations relatives aux différents petits domaines (ex : box-plot).

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…

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

- 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 aspects à prendre en compte.

- La qualité 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.

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

- En terme technique comme en interprétation, la complexité est surtout apportée par les modèles qualitatifs avec effet aléatoire : le L**M**M n'est jamais qu'un modèle linéaire "classique" avec une structure de variance-covariance un peu plus compliquée et le GLM a une densité qui reste mathématiquement gérable

****