Méthode d'estimation sur petits domaines pour l'indicateur AROPE au niveau régional

Olivier Sautory
DMCSI - Insee

Journées de méthodologie statistique 2018





Insee
Mesurer pour comprendre

Plan de la présentation

- Le contexte : réglementation européenne sur la précision d'indicateurs régionaux de pauvreté
- Quelques éléments de théorie des sondages : estimation par régression, calage
- Les données : enquête SRCV et variables auxiliaires
- Estimation de totaux régionaux par une méthode « petits domaines » : estimateur « synthétique »
- Application aux données SRCV (résultats provisoires)
- Reste à faire

Le contexte

Futur règlement européen sur les statistiques sociales (Integrated European Social Statistics, ou IESS)

→ contraintes en matière de précision pour 12 indicateurs, dont 2 définis au niveau régional (nomenclature « NUTS 2 » = anciennes régions) :

- estimation trimestrielle du nombre de chômeurs, à partir de l'enquête emploi en continu,
- indicateur de pauvreté AROPE (at-risk of poverty or social exclusion), à partir du dispositif Statistique sur les revenus et les conditions de vie (SRCV).

Problème : il faudrait augmenter significativement la taille de l'échantillon actuel de SRCV pour respecter les contraintes de précision sur l'AROPE.

→ l'utilisation de méthodes d'estimation sur petits domaines est privilégiée.

Le contexte

Des travaux menés par Pascal Ardilly sur le sujet :

Regional estimates of poverty indicators based on a calibration technique, Statistical working papers, Eurostat, 2015

Principe : en réponse à Eurostat, qui exige des poids individuels, méthode d'estimation, dite « synthétique » = calage de l'échantillon national de SRCV sur des structures régionales.

Objet de la communication : présentation de résultats de travaux poursuivant ceux de P. Ardilly.

Données utilisées : échantillon SRCV 2010

Théorie : notations, données

Population $U = \{1 \dots N\}$

Échantillon $s = \{1 \dots k \dots n\}$ tiré dans U selon un plan de sondage p.

Variable d'intérêt $\boldsymbol{Y}$: total $Y = \sum_{k \in U} y_k$ (à estimer)

Variables auxiliaires $\boldsymbol{X_1 \dots X_j \dots X_J}$ (dont la constante $\boldsymbol{X_1} = 1$), connues sur s, et dont les totaux X_j sur U sont connus.

$$x_k = \begin{pmatrix} 1 \\ x_{jk} \\ x_{Jk} \end{pmatrix}, \quad X_j = \sum_{k \in U} x_{jk}, \quad X = \sum_{k \in U} x_k = \begin{pmatrix} N \\ X_j \\ X_J \end{pmatrix}$$

Poids d_k conduisant à des estimateurs (approximativement) sans biais :

$$\hat{Y} = \sum_{k \in s} d_k \, y_k \qquad \hat{X}_j = \sum_{k \in s} d_k \, x_{jk} \quad j = 1 \dots J$$

Estimation par régression

On suppose qu'il existe une relation linéaire approchée (voire très approchée) entre $\boldsymbol{Y}$ et les variables auxiliaires $\boldsymbol{X_j}$:

$$\forall k \in U \qquad y_k = \sum_{j=1}^{J} b_j x_{jk} + \varepsilon_k = {}^t b\, x_k + \varepsilon_k \quad , \text{ avec } {}^t b = \left(b_1 ... b_j ... b_J\right)$$

- b est estimé **dans la population** par les m.c.o. $B = \left(\sum_{k \in U} x_k {}^t x_k\right)^{-1} \left(\sum_{k \in U} x_k y_k\right)$

On a : $Y = \sum_{k \in U} y_k = \sum_{k \in U} {}^t B\, x_k$ (car terme constant $\rightarrow$ somme des résidus nulle)

soit :

$$Y = {}^t B\, X$$

- B inconnu, estimé (approximativement) sans biais **dans l'échantillon** par :

$$\hat{B} = \left(\sum_{k \in s} d_k\, x_k {}^t x_k\right)^{-1} \left(\sum_{k \in s} d_k\, x_k\, y_k\right) \qquad \text{(régression pondérée dans s)}$$

Estimation par régression

Estimateur par régression : $\hat{Y}^r = {}^t\hat{B}\,X$

- on montre que : $\hat{Y}^r = \sum_{k \in s} w_k\, y_k$ où les poids w_k ne dépendent que des x_k

  et, pour toute variable auxiliaire X_j : $\hat{X}_j^r = \sum_{k \in s} w_k\, x_{jk} = X_j$

  i.e. l'échantillon est **calé sur les totaux X_j**.

- $\hat{Y}^r$ approximativement **sans biais**

- la variance de $\hat{Y}^r$ est fonction des résidus de la régression de Y sur X_1 … X_j … X_J : plus les résidus sont petits, plus faible est la variance

→ $\hat{Y}^r$ meilleur que $\hat{Y}$ si la relation linéaire est « bonne »

Estimation par calage

On cherche des poids w_k proches des poids initiaux d_k qui assurent le calage de l'échantillon sur les totaux X_j :

$$\min_{w_k} \sum_{k \in s} D(d_k, w_k) \text{ avec } \sum_{k \in s} w_k x_k = \sum_{k \in U} x_k = X$$

Si on choisit $D(d_k, w_k) = \dfrac{(w_k - d_k)^2}{d_k}$, méthode « linéaire » (M = 1) de la macro Calmar, **les poids w_k sont ceux de l'estimateur par régression**

→ ces poids peuvent être récupérés en sortie de Calmar, et on a, pour toute variable d'intérêt ***Y*** (quantitative ou qualitative) :

$$\sum_{k \in s} w_k y_k = \hat{Y}^r = {}^t\hat{B} X$$

Les données : enquête SRCV et variables auxiliaires

6 indicateurs de pauvreté sont calculés au niveau régional (demande Eurostat), dont :

1. le taux de risque de pauvreté (*at risk of poverty rate)* : part des individus ayant un niveau de vie inférieur à 60% du niveau de vie médian national (appelé seuil de risque de pauvreté)
2. l'indicateur AROPE (*at risk of poverty or social exclusion rate*) : part des individus étant dans au moins l'une des situations suivantes :
   - avoir un revenu disponible inférieur au seuil de risque de pauvreté
   - être dans un état de « privation matérielle aigüe »
   - vivre dans un ménage à faible intensité de travail.

→ estimer un effectif d'individus « pauvres » $Y = \sum_{k \in U} y_k$

U = population de ménages, y_k nombre d'individus « pauvres » du ménage k.

Les données : enquête SRCV et variables auxiliaires

On mobilise des variables auxiliaires « explicatives » de la pauvreté :

- connues pour tout ménage k de l'échantillon s_{NAT}
- dont on connaît les totaux X_{NAT} sur la population nationale U_{NAT}
- dont on connaît les totaux X_{REG} sur la population régionale U_{REG} (pour toute région REG).

- Variables issues du RP : sexe, 6 tranches d'âge, 4 niveaux de diplôme, nationalité (5 modalités), 11 catégories sociales, appartenance à une ZUS, 3 tranches d'unité urbaine, 5 types de ménage, locataire HLM

- Fichier Revenus Disponibles Localisés (RDL) : distribution des niveaux de vie (revenu disponible par unité de consommation) – 20 modalités

- Nombre de bénéficiaires de l'Allocation de Solidarité aux Personnes Âgées (ASPA)

Les variables individuelles sont agrégées au niveau ménage.

Les « régressions » nationale et régionales

Hypothèse : $\boldsymbol{Y}$ (nombre de ménages pauvres) fonction des variables « explicatives » de la pauvreté $\boldsymbol{X_1} \ldots \boldsymbol{X_j} \ldots \boldsymbol{X_J}$ (dont la constante)

Au niveau national :

$$\forall k \in U_{NAT} \quad y_k = {}^t b_{NAT}\, x_k + \varepsilon_k \rightarrow B_{NAT}\ (\text{m.c.o. sur } U_{NAT})$$

$$\rightarrow \hat{B}_{NAT} = \left(\sum_{k \in s_{NAT}} d_k\, x_k\, {}^t x_k \right)^{-1} \left(\sum_{k \in s_{NAT}} d_k\, x_k\, y_k \right) \text{ (régression pondérée sur } s_{NAT})$$

Pour une région donnée REG :

$$\forall k \in U_{REG} \quad y_k = {}^t b_{REG}\, x_k + \varepsilon_k \rightarrow B_{REG}\ (\text{m.c.o. sur } U_{REG})$$

$$\rightarrow \hat{B}_{REG} = \left(\sum_{k \in s_{REG}} d_k\, x_k\, {}^t x_k \right)^{-1} \left(\sum_{k \in s_{REG}} d_k\, x_k\, y_k \right) \text{ (régression pondérée sur } s_{REG})$$

Estimation d'un total régional : basique

On cherche à estimer un total régional $Y_{REG} = \sum_{k \in U_{REG}} y_k$

0. Estimateur « basique »

Utilisation des poids d_k :

$$\hat{Y}^{bas}_{REG} = \sum_{k \in s_{REG}} d_k \, y_k = \sum_{k \in s_{NAT}} d_k \, y_k \, \mathbb{1}_{s_{REG}}(k) \quad \text{où} \quad s_{REG} = s_{NAT} \cap U_{REG}$$

- (approximativement) sans biais,
- de variance dépendant de n_{REG} = card (s_{REG})

Estimation d'un total régional : par régression

1. Estimateur par régression - calage national

Calage national (s_{NAT}) sur les marges nationales X_{NAT} (méthode linéaire)

$$\sum_{k \in s_{NAT}} w_k^{NAT} x_{jk} = X_{j\,NAT} \quad \forall\ j = 1...J \rightarrow \text{poids } w_k^{NAT}$$

$\rightarrow$ estimateur par régression : $\hat{Y}^r_{REG} = \sum_{k \in s_{REG}} w_k^{NAT} y_k = \sum_{k \in s} w_k^{NAT} y_k \mathbb{1}_{s_{REG}}(k)$

Approximativement sans biais, de variance dépendant de n_{REG} et des résidus de la régression dans s de la variable $\boldsymbol{Y}\ \boldsymbol{\mathbb{1}}_{s_{REG}}$ sur les variables auxiliaires $\boldsymbol{X_1} \ldots \boldsymbol{X_j} \ldots \boldsymbol{X_J}$.

En général : $V\left(\hat{Y}^r_{REG}\right) < V\left(\hat{Y}^{bas}_{REG}\right)$

Estimation d'un total régional : par régression spécifique

2. Estimateur par régression spécifique à la région - calage régional

Calage régional (s_{REG}) sur les marges régionales X_{REG} (méthode linéaire)

$$\sum_{k \in s_{REG}} w_k^{REG} x_{jk} = X_{j\,REG} \quad \forall\ j = 1...J \rightarrow \text{poids } w_k^{REG}$$

→ estimateur par régression spécifique : $\hat{Y}_{REG}^{r/REG} = \sum_{k \in s_{REG}} w_k^{REG} y_k = {}^t\hat{B}_{REG} X_{REG}$

Approximativement sans biais, de variance dépendant de n_{REG} et des résidus de la régression dans s_{REG} de la variable $\boldsymbol{Y}$ sur les variables auxiliaires $\boldsymbol{X_1}$ … $\boldsymbol{X_j}$ … $\boldsymbol{X_J}$.

En général : $V\left(\hat{Y}_{REG}^{r/REG}\right) < V\left(\hat{Y}_{REG}^{r}\right)$

Estimation d'un total régional : petits domaines (synthétique)

3. Estimateur synthétique

Modèle (M) : $\mathbf{B_{NAT} = B_{REG}}$ pour toute région REG

d'où : $Y_{REG} = {}^{t}B_{REG}\, X_{REG} \overset{(M)}{=} {}^{t}B_{NAT}\, X_{REG}$

D'où l'estimateur synthétique : $\boxed{\hat{\mathbf{Y}}^{\mathbf{syn}}_{\mathbf{REG}} \overset{\mathbf{(def)}}{=} {}^{\mathbf{t}}\hat{\mathbf{B}}_{\mathbf{NAT}}\, \mathbf{X}_{\mathbf{REG}}}$

On montre que : $\hat{Y}^{syn}_{REG} = \sum_{k \in s_{NAT}} w^{REG}_{k/syn}\, y_k$ où $w^{REG}_{k/syn}$ = poids obtenus par **calage national (s_{NAT}) sur les marges régionales X_{REG}** (méthode linéaire)

$$\sum_{k \in s_{NAT}} w^{REG}_{k/syn}\, x_{jk} = X_{j\,REG} \quad \forall\, j=1...J$$

Estimation d'un total régional : petits domaines (synthétique)

Variance : $$V\left(\hat{Y}_{REG}^{syn}\right) = V\left({}^{t}\hat{B}_{NAT}\, X_{REG}\right) = {}^{t}X_{REG}\, V\left(\hat{B}_{NAT}\right) X_{REG}$$

dépendant de n = card (s_{NAT}), donc en général (nettement) inférieure aux variances des estimateurs précédents. MAIS… **biaisé** !

$$\begin{aligned} Biais\left(\hat{Y}_{REG}^{syn}\right) &= E\left(\hat{Y}_{REG}^{syn}\right) - Y_{REG} = E\left({}^{t}\hat{B}_{NAT}\, X_{REG}\right) - Y_{REG} \\ &\cong {}^{t}B_{NAT}\, X_{REG} - {}^{t}B_{REG}\, X_{REG} = {}^{t}\left(B_{NAT} - B_{REG}\right) X_{REG} \end{aligned}$$

Biais petit si :

coeff de régression dans U_{NAT}, $\mathbf{B_{NAT} \approx B_{REG}}$, coeff de régression dans U_{REG}

Erreur quadratique moyenne :

$$EQM\left(\hat{Y}_{REG}^{syn}\right) = E\left(\hat{Y}_{REG}^{syn} - Y_{REG}\right)^2 = V\left(\hat{Y}_{REG}^{syn}\right) + Biais^2\left(\hat{Y}_{REG}^{syn}\right)$$

Estimation de l'EQM de l'estimateur synthétique

Si $\hat{Y}_{REG}$ désigne un estimateur (approximativement) sans biais de Y_{REG}, on montre que l'erreur quadratique moyenne s'écrit :

$$EQM\left(\hat{Y}_{REG}^{syn}\right) = E\left(\hat{Y}_{REG}^{syn} - \hat{Y}_{REG}\right)^2 - V\left(\hat{Y}_{REG}^{syn} - \hat{Y}_{REG}\right) + V\left(\hat{Y}_{REG}^{syn}\right)$$

estimée (approximativement) sans biais par :

$$\hat{EQM}\left(\hat{Y}_{REG}^{syn}\right) = \left(\hat{Y}_{REG}^{syn} - \hat{Y}_{REG}\right)^2 - \hat{V}\left(\hat{Y}_{REG}^{syn} - \hat{Y}_{REG}\right) + \hat{V}\left(\hat{Y}_{REG}^{syn}\right)$$

instable, pouvant prendre des valeurs négatives.

Dans la suite, on calculera un majorant de cette EQM estimée :

$$\left(\hat{Y}_{REG}^{syn} - \hat{Y}_{REG}\right)^2 + \hat{V}\left(\hat{Y}_{REG}^{syn}\right)$$

et on prendra l'estimateur par régression spécifique (calage régional) comme estimateur $\hat{Y}_{REG}$

Application aux données SRCV

- Non prise en compte du plan de sondage (très complexe…) de l'enquête SRCV
- Mais prise en compte des poids inégaux
- Utilisation de la procédure SURVEYREG de SAS pour estimer la matrice de variance de $\hat{B}_{NAT}$

- Calcul des gains de précision par rapport à l'estimateur de référence, qui est l'estimateur par régression (national) :

Estimateur par régression spécifique : $\dfrac{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r/REG}\right)}{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r}\right)}\times 100$

Estimateur synthétique :

$$\frac{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{SYN}\right)}{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r}\right)}\times 100 \quad \text{et} \quad \frac{\sqrt{\hat{\mathrm{EQM}}\left(\hat{Y}_{REG}^{SYN}\right)}}{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r}\right)}\times 100$$

Taux de risque de pauvreté

Région	nobs	Taux de pauvreté (%)				Gains de précision (%)		
		cal-natio	cal-regio	synth	CV-synth	Ecarts-types		R(EQM)(3)/(1)
		(1)	(2)	(3)	%	(2)/(1)	(3)/(1)	
11	1729	10,5	12,3	13,0	3,08	53	38	79
21	288	16,7	16,1	15,0	1,50	51	6	33
22	409	22,3	15,5	14,9	1,65	17	7	17
23	286	13,9	15,1	13,5	1,36	21	3	25
24	416	10,7	11,5	12,2	1,69	28	7	24
25	271	8,8	13,1	13,6	1,88	83	13	25
26	321	13,0	13,8	12,8	1,90	36	10	42
31	788	20,9	18,4	18,9	1,53	31	12	23
41	483	19,9	14,1	14,2	1,73	36	8	9
42	297	11,2	12,0	11,7	2,11	24	10	17
43	261	15,9	11,8	13,0	1,81	31	6	32
52	775	10,2	11,7	11,5	1,81	39	13	16
53	628	14,1	11,7	11,5	1,94	37	10	13
54	361	17,5	13,8	13,8	1,83	40	7	8
72	679	16,9	12,1	13,0	1,94	22	8	31
73	512	14,9	14,1	14,1	1,84	29	12	12
74	166	18,0	16,5	14,6	1,79	36	6	39
82	907	9,2	12,4	12,2	1,78	42	18	27
83	249	14,1	13,0	13,9	1,76	25	7	27
91	444	19,3	19,8	18,7	1,53	21	6	25
93	739	12,9	16,3	15,7	1,95	51	17	33
94	35	19,6	23,3	18,9	2,17	51	5	55

Indicateur AROPE

Région	nobs	Taux de pauvreté (%)			CV-synth	Gains de précision (%)		
		cal-natio	cal-regio	synth		Ecarts-types		R(EQM)(3)/(1)
		(1)	(2)	(3)	%	(2)/(1)	(3)/(1)	
11	1729	16,4	18,4	18,1	3,13	74	43	47
21	288	24,5	21,9	21,4	1,73	48	9	15
22	409	29,5	20,0	21,0	1,95	30	10	27
23	286	17,7	19,8	19,6	1,71	26	5	6
24	416	16,5	16,7	17,9	1,89	36	11	39
25	271	15,5	17,4	19,7	2,06	73	14	80
26	321	15,2	16,1	18,7	2,01	50	15	104
31	788	29,3	25,4	25,5	1,61	36	14	15
41	483	28,6	21,4	20,2	1,82	44	10	35
42	297	17,4	18,7	17,0	2,08	49	11	58
43	261	18,9	16,3	18,8	2,01	43	9	61
52	775	14,0	15,3	17,1	2,02	36	20	108
53	628	19,1	16,5	17,1	2,08	42	14	28
54	361	23,2	21,5	19,7	1,95	45	9	44
72	679	23,2	17,5	18,7	1,95	36	11	38
73	512	20,5	20,0	19,5	1,87	49	15	27
74	166	25,7	20,7	20,6	1,79	39	7	7
82	907	13,0	16,8	17,5	1,86	61	24	53
83	249	19,0	18,6	19,8	1,83	43	9	32
91	444	25,9	27,1	24,5	1,56	40	8	53
93	739	17,6	22,3	21,4	1,97	59	22	49
94	35	28,8	26,5	24,9	2,35	113	6	18

Reste à faire…

- Travailler sur des données plus récentes

- Etudier la possibilité de prendre en compte de nouvelles variables auxiliaires

- Mettre en œuvre des techniques permettant d'apprécier le biais

- Améliorer l'estimation de l'EQM de l'estimateur synthétique

- Prendre en compte le plan de sondage dans les calculs de précision

- Comparer avec des méthodes d'estimation sur petits domaines reposant sur des modélisations explicites (ex : Fay-Herriot, Battese-Harter-Fuller)

Merci pour votre attention !