Méthode d'utilisation d'enquête à un niveau géographique où l'échantillon est faible

François JEGER
Insee

I - LA METHODE UTILISEE

La méthode utilisée pour l'estimation du revenu moyen par ménage pour les bassins d'emploi d'Alsace est celle présentée par Robert FAY et Roger HERRIOT dans "Journal of the Ameran Statistical Association" Juin 1979 sous le titre "Estimates of Income for Small Places : "An Application of James STEIN Procedures to Census Date".

L'idée est la suivante :

le revenu par ménage est connu par l'enquête revenus fiscaux avec des erreurs d'échantillonnage fortes dans les zones géographiques fines. D'autres variables, dites auxiliaires, correlées avec ce revenu, sont mieux connues à partir de sources exhaustives : par exemple, le revenu par foyer fiscal.

Une régression où le revenu par ménage est expliqué par ces variables auxiliaires fournit des "revenus théoriques" pour chacune de ces zones.

Le revenu par ménage est alors estimé par une moyenne de la valeur observée et de la valeur théorique pondérées par l'inverse de leurs variances respectives. Autrement dit, si l'échantillon dans une zone est faible on tiendra peu compte de la valeur observée et beaucoup de la valeur théorique, si l'échantillon est fort ce sera l'inverse.

a - La formulation mathématique : le James Stein et ses dérivés

Le James-Stein - ou estimateur raccourci - est un estimateur biaisé d'un paramètre dont la variance est plus faible que la moyenne d'un échantillon. Le risque quadratique peut être ainsi réduit. Herriot en propose une extension sous la forme suivante :

Soit $Yi \sim N(\Theta i, Di)$ le revenu moyen observé dans la zone i, Oi étant inconnu (à estimer), Di la variance liée à la taille de l'échantillon de la zone i.

On suppose par ailleurs que $\Theta i \sim N(Xi\,\beta, A)$ Xi étant une ligne de variables auxiliaires et A la variance inconnue de l'explication de Oi par les variables auxiliaires.

Si on suppose que les Yi - Oi (erreurs d'échantillonnage) et les $\Theta i - Xi$ (erreurs "d'explication" des Oi par les Xi) sont indépendants.

$Y \sim N(X\,\beta, A + D)$ où A + D diagonale de terme A + Di

L'estimation des moindres carrés généralisés fournit un revenu théorique

$$(1) \qquad Yi^{*} = Xi\,\hat{\beta} = Xi\,(X^{t}\,(A + D)^{-1}\,X)^{-1}\,X\,(A + D)^{-1}\,Y$$

L'estimateur retenu est

$$\overset{*}{\Theta i} = \frac{A\,Yi + Di\,Yi}{A + Di}$$

A est alors déterminé par les résidus, puisque

$$E\left(\sum_{i=1}^{n} \frac{(yi - \overset{*}{yi})^2}{A + Di}\right) = K - p$$

où k est le nombre d'observations
p est le nombre de paramètres explicatifs
yi le revenu observé
yi* le revenu théorique

Pour des raisons asymptotiques, A sera défini par l'égalité

$$(2) \qquad \sum_{i=1}^{n} \frac{(yi - yi^{*})^2}{A + Di} = k - p$$

Il faudrait pour calculer yi dans (1) connaître la valeur de A donnée par l'équation (2) qui suppose yi* déjà connu, pour rompre ce cercle vicieux une méthode itérative est possible.

b - Résolution itérative des équations 1 et 2

La valeur Yi* dépend de la variance A, posons Yi* Yi* (An) où An est un estimateur convergent de A.

Chaque régression donne une série de résidus Yi - Yi* (An) où An est l'estimation de A détenue par la précédente régression.

Ces résidus vont permettre de définir un meilleur estimateur A de A de la façon suivante :

en posant $f(An)) = \frac{(Yi - Yi^{*}(An))^2}{An + Di}$

$g(An) = \frac{(Yi - Yi^{*}(An))^2}{(An + Di)^2}$

A_{n+1} = An + (k - p - f (An))/ g (A n) (si cette valeur est positive, sinon A_{n+1} = 0)

La suite An converge vers une valeur A telle que k - p = f (A) = O (c'est-à-dire l'équation 2) et on initialise par Ao = 0

Remarque : la convergence est rapide (moins de 10 itérations) elle se justifie par l'approximation par g de la dérivée de f par rapport à A.

f' (An) (An + 1 - An) + f (An) = k - p définit un algorithme convergent si f est contractante.

II - APPLICATION A L'EVALUATION DU REVENU MOYEN PAR MENAGE PAR BASSIN D'EMPLOI EN ALSACE

Le revenu moyen par ménage est connu par l'enquête revenus fiscaux de 1979 avec un faible échantillonnage au niveau du bassin d'emploi (une dizaine d'observations dans certains bassins d'emploi).

Soit Yi ce revenu moyen observé et Di l'erreur d'échantillonnage associée.

Les variables auxiliaires sont tirées de la banque de données locales.

X_i^1 = revenu imposable par foyer fiscal (source exhaustive)

X_i^2 = nombre moyen de foyers fiscaux par ménage

X_i^3 = taille moyenne du ménage

X_i^4 = % d'agriculteurs dans le bassin d'emploi

X_i^5 = % d'étrangers dans le bassin d'emploi

X_i^6 = % de la population âgée de + de 60 ans

$Yi \sim N(\Theta i, Di)$ Θi à estimer

$\Theta i = Xi\,\beta + ui$ avec $ui \sim N(O,A)$ A à estimer.

La méthode présente conduit à calculer un revenu théorique

$$Yi^* = Xi\ (Xt\ (A+D)^{-1}\ X)^{-1}\ Xt\ (A+D)^{-1}\ Y$$

ou A + D est la matrice diagonale de terme général A + Di dont on calcule la moyenne avec Yi pondérée par l'inverse des variances

$$\theta i^* = \frac{Yi^*\ Di + Yi\ A}{A + Di}$$ est l'estimateur du revenu moyen par ménage

Il est possible de transformer la régression des moindres carrés généralisés en une régression des moindres carrés ordinaires pondérés. En effet puisque la matrice A + D est diagonale en posant :

$$\sigma\ i = \frac{1}{\sqrt{A + Di}}$$

la régression devient :

$$Y = X\beta + u \quad \text{avec } u \sim N(0,\ \sigma^{-2})$$

où σ est la matrice diagonale donc le terme général est σ_i soit encore

$$\sigma\ Y = \sigma\ X\ \beta + \upsilon \text{ avec } \upsilon \sim N(0,I)$$

L'estimateur de β est donc le même que celui de la régression des moindres carrés ordinaires où les observations sont pondérées par les Ti.

La procédure itérative permettant de déterminer à la fois A et Y_i^* (décrite en b) est donc simplifiée puisque :

$$f\ (A) = \frac{\sum \hat{u}\ 2_i}{A + Di} = \sum (\sigma i\ \hat{u}^2_i) = \sum \hat{\upsilon}_i^2$$

et

$$g\ (A) = -\frac{\sum \hat{u}\ 2_i}{(A + Di)^2} = -\sum (\sigma^2 i\ \hat{u}_i^2) = -\sum \sigma_i\ \hat{\upsilon}_i^2$$

Le passage d'une itération à une suivante est donc définie par

$$A_{n+1} = An + (k - p - f\ (An))/\ g(\ An)).$$

III - LES RESULTATS

a - Régression sur 90 départements français

Le revenu moyen par ménage (REME) est expliqué dans un modèle linéaire en fonction du revenu moyen par foyer fiscal (RF84),

du nombre de foyers par ménage (FOYME), du % d'agriculteurs (PAGRI), le % d'étrangers (PETR), le % de la population de + 60 ans (PAGE), la taille des ménages (TAMEM).

Sur 90 observations (départements) pondérées par le carré du nombre d'observations de l'enquête revenu 79 (pour tenir compte de la variance d'échantillonnage) le R2 est 0,95.

La régression est alors :

REME = - 6.799 + 0,769 RF84 + 45.407 FOYME + 8.509 PETR
(- 0,47) (13,45) (4,5) (0,626)

- 11.449 TAMEM - 57.510 PAGRI - 7.808 PAGE + u
(- 2,9) (- 2,25) (- 0,42) (R2 = 0,95)

Les chiffres entre parenthèses sont les students de chaque paramètre.

Les résidus suivent une loi normale avec une probabilité de 0,85, les "students" de PETR et PAGE n'étant pas significativement différents de 0, la proportion d'agriculteurs n'étant pas un indicateur démographique et pouvant avoir un effet en Alsace opposé à celui de la France entière, nous retiendrons 3 variables auxiliaires :

la régression devient alors :

REME =
- 24.850 + 0,89 RF84 + 51.530 FOYME - 11.303 TAMEN + u
(- 2,9) (16,1) (4,8) (3,7) R2 = 0,94

La taille du ménage étant correlée au nombre de foyers par ménage, nous retiendrons seulement deux variables auxiliaires : le revenu par foyer fiscal et le nombre de foyers fiscaux par ménage, la régression est alors :

REME = - 23.037 + 1,054 RF84 + 16.995 FOYME + u
(- 2,5) (28,6) (2,92) R2 = 0,93

Le R2 diminue peu d'une régression à l'autre, les deux dernières variables auxiliaires sont donc suffisantes à l'explication de REME. De plus, dans la dernière équation le coefficient de RF84 est vraisemblable (si on admet un coefficient d'actualisation 1,35 entre 79 et 84 et que l'abattement des revenus imposables est 0,3).

Nous calculerons sur la base de la dernière régression les estimations de revenus de bassin d'emploi en Alsace. Les estimations résultant des autres régressions sont présentées en annexe.

b - Estimations des revenus des bassins d'emploi

Si THEO est le revenu théorique d'un bassin d'emploi, c'est-à-dire calculé à partir de la dernière régression et VARTH sa

variance associée (mesure de l'erreur de la régression), si RMEEF est le revenu observé dans l'enquête revenu (faible échantillon) et VAREMP sa variance associée (mesure de l'erreur de l'échantillonnage), la méthode d'HERRIOT conduit à retenir l'estimateur du revenu moyen.

$$ESTIM = \frac{THEO \times VAREMP + RMEEF \times VARTH}{VAREMP + VARTH}$$

(moyenne pondérée des revenus théoriques et observés par l'inverse de leur variance)

dont le coefficient de variation est :

$$CVEST = \left(\frac{VAREMP \times VARTH}{VARTH + VARTH} \right)^{1/2}$$

Le tableau suivant donne ces estimations de revenu moyen par ménage par bassin d'emploi en Alsace. On trouvera aussi en annexe les estimations résultant d'un choix de régression différent.

Estimation des revenus moyens par ménage par bassin d'emploi en Alsace

SAS 10:29 WEDNESDAY, SEPTEMBE

OBS	NOM	THEO	VARTH	RMEEF	VAREMP	ESTIM	CVEST	INF	SUP
1	WISSEMBO	62525.3	1622609	70121	83096533	62543.3	1251.50	60070.0	65015.9
2	NIEDERBR	62075.3	1242798	59717	43150053	62009.3	1099.09	59855.0	64163.5
3	HAGUENAU	65134.1	1260762	67266	61684374	66557.3	1111.53	64378.7	68736.0
4	SARRE-UN	54292.3	903357	64704	241148599	54331.4	951.55	52466.4	56196.5
5	SAVERNE	60261.1	986283	63615	40468632	60340.9	981.23	58417.7	62264.1
6	STRASBOU	63930.4	125690	78641	39530204	69011.0	353.97	68317.2	69704.8
7	MOLSHEIM	67917.9	994934	72094	89306658	67970.5	991.98	66026.2	69914.8
8	SCHIRMEC	55345.2	3710+5	47737	82437865	55309.4	607.77	54116.1	56500.6
9	SELESTAT	61636.1	455004	63824	79340658	61648.5	672.63	60330.2	62966.8
10	STE-MARI	55335.2	833504	47035	107302666	55313.4	531.75	54271.2	56355.7
11	COLMAR	66444.5	395488	78872	50338413	66541.3	626.42	65313.5	67769.1
12	RF-BRISS	69369.4	1113290	75807	97119251	69473.9	1051.46	66413.0	70534.7
13	GUEBWILL	64570.5	900618	73888	54594365	64705.2	888.23	62964.2	66446.3
14	THANN-CE	62739.7	677520	60138	29294290	62680.8	813.94	61085.5	64276.2
15	MULHOUSE	67197.5	157511	72213	13036793	67257.4	394.50	66484.2	68030.6
16	ST-LOUIS	76155.7	295234	65540	51975448	76096.0	541.82	75034.0	77159.0
17	ALTKIRCH	66535.3	1153996	59055	168794662	66285.9	1070.59	64187.5	68384.3

369

↑ Bassin d'emploi

↑ Revenu par ménage théorique expliqué par:
- revenu par foyer
- nombre de foyers par ménage

Variance Theorique

↑ Revenu observé (faible échantillon)

Variance empirique

↑ ESTIMATION (moyenne pondérée de THEO et RMEEF) des revenus moyens par ménage en 79

Coefficient de variation de ESTIM

Intervalle de confiance à 95%

c - Régression sur 34 bassins d'emploi. Calcul de la variance théorique par itération

Pour déterminer un revenu théorique fonction du revenu par foyer fiscal et des autres variables auxiliaires, la méthode d'Herriot est appliquée sur 34 bassins d'emploi d'Alsace et Lorraine. La variance théorique est calculée suivant la méthode décrite en I.b. Cinq itérations suffisent pour obtenir la valeur de la variance théorique A.

Nbre d'itérations	R^2 de la régression sur 34 bassins d'emploi REME = f(RF84,PAGE,PAGRI,TAMEN)	A	$\left\lvert k-p\frac{\sum(\hat{y}i - yi)^2}{A + Di}\right\rvert$
1	0,445	8.10^6	12
2	0,442	13.10^6	4,09
3	0,440	$14,5.10^6$	0,68
4	0,440	$14,57.10^6$	0,04
5	0,440	$14,57.10^6$	0,002

Le R2 de la régression ne s'améliore pas d'une itération à l'autre, par contre l'égalité

$$\sum \frac{(yi - yi)2}{A + Di} = 34 - 4$$

est rapidement vérifiée avec une variance théorique de 14,57.10 6. Cette valeur est supérieure à celle issue de la régression sur 90 départements. Cela provient en partie du faible nombre de bassins d'emplois.

Les régressions sont successivement :

itération 1 : REME = - 61.282 + 1,097 RF84 - 169 PAGE + 302 PAGRI + 20.267 TAMEM

itération 2 : REME = - 58.431 + 1,091 RF84 - 213 PAGE + 300 PAGRI + 20.505 TAMEM

itération 3 : REME = - 57.171 + 1,08 RF84 - 232 PAGE + 298 PAGRI + 20.267 TAMEM

itération 4 : REME = - 56.924 + 1,087 RF84 - 236 PAGE + 298 PAGRI + 20.221 TAMEM

itération 5 : REME = - 56.907 + 1,087 RF84 - 236 PAGE + 298 PAGRI + 20.218 TAMEM

Les estimations de revenus par bassin d'emploi d'Alsace et de Lorraine sont présentées à la page suivante.

En ajoutant à ces 34 observations, 13 bassins d'emploi de Franche-Comté, on n'améliore pas le R2 de la régression (0,40) et la variance théorique est 17.10 6.

ANNEXE

Estimation des revenus moyens par ménages en 79. Régression sur les 34 bassins d'Alsace et de Lorraine. Variance théorique calculée par situation

	OBS	DE	THEO	VARTH	REME	VAREMP	ESTIM	CVEST	INF	SUP
Lorraine	1	4111	55510.2	14579517	53334.2	17936618	53269.1	2836.93	50708.3	61829.6
	2	4112	53229.9	14579517	57682.5	40260013	54413.7	3871.03	49001.3	60926.1
	3	4113	53747.6	14579617	61590.2	10157235	61836.7	2446.74	57091.0	66582.3
	4	4121	54095.7	14579617	49045.7	25295185	52256.4	3041.14	46295.8	58217.1
	5	4122	55761.0	14579617	59120.7	16931563	64300.5	2793.91	59314.7	69784.5
	6	4123	63904.0	14579617	64731.5	37449234	64410.4	3769.39	56996.1	71832.7
	7	4130	53442.4	14579617	61395.5	11762813	65649.1	2551.53	60643.1	70655.1
	8	4141	55029.9	14579617	59630.9	29045944	67233.3	3115.63	61120.7	73339.9
	9	4142	65107.9	14579617	59326.0	54350123	61932.5	3139.35	55711.6	68253.3
	10	4150	53901.5	14579617	61372.9	20403684	59673.1	2916.06	54157.7	65588.6
	11	4160	52935.2	14579617	44000.1	27162930	50512.9	3080.15	44475.8	56550.0
	12	4171	60256.9	14579617	56200.8	49149463	61624.4	3353.23	55052.1	68196.7
	13	4172	52043.9	14579617	52541.2	37732396	52197.1	3243.47	45839.9	58554.3
	14	4180	60081.6	14579617	57050.3	37752575	59237.1	3243.11	52880.6	65593.6
	15	4191	56643.4	14579617	52745.1	12229006	54523.3	2576.39	49463.7	59577.9
	16	4192	62146.6	14579617	63045.3	138036363	64647.5	3031.36	56930.1	71165.0
	17	4193	54725.7	14579617	65185.9	24990024	57643.7	3034.19	51890.6	63790.7
Alsace	18	4271	67153.9	14579617	70121.9	93359761	67554.8	3551.13	60594.6	74515.0
	19	4272	65909.0	14579617	59717.6	42837712	64262.2	3298.11	57797.9	70726.5
	20	4273	70292.3	14579617	67266.6	67004166	73326.2	3460.37	66543.8	80108.5
	21	4274	61063.4	14579617	64704.4	223597793	61236.3	3099.61	54035.1	68537.6
	22	4275	65561.1	14579617	63615.8	33708604	62364.2	3254.33	56485.7	69242.7
	23	4276	67095.6	14579617	79641.4	43637672	69987.1	3305.21	63507.7	76466.4
	24	4277	70571.8	14579617	72694.9	33631346	70972.1	3523.53	64066.0	77878.2
	25	4278	53270.7	14579617	47737.8	74155698	52369.9	3490.58	45528.3	59211.4
	26	4279	63373.9	14579617	63824.9	70371267	63450.9	3477.36	56615.3	70266.5
	27	4281	50497.0	14579617	47095.4	106622809	50080.6	3531.31	43067.2	57106.0
	28	4282	64305.6	14579617	73872.3	43174219	68073.7	3345.49	61516.6	74630.9
	29	4283	79217.4	14579617	75807.7	95040433	78763.9	3555.35	71795.4	85732.4
	30	4284	66005.0	14579617	73338.0	52005342	67731.1	3374.50	61117.1	74345.1
	31	4285	64535.6	14579617	60138.6	25216183	62924.7	3039.44	56967.4	68882.0
	32	4286	68814.6	14579617	72213.2	13039072	70605.4	2826.23	65453.0	75752.9
	33	4287	70267.7	14579617	65540.5	50661572	73870.5	3364.74	67275.6	80465.4
	34	4288	70336.0	14579617	59056.0	149584929	69334.2	3644.53	62190.3	76479.0

ANNEXE

Estimation basée sur la régression : 90 départements, 6 variables (RF84, TAMEM, PAGRI, PETR, PAGE et FOYME)

OBS	Ncc	THEO	VARTH	RMEEF	VAREMP	ESTIM	CVEST	INF	SUP
1	WISSEMBO	64346.5	2634779	70121	83096533	64524.0	1528.03	61391.3	67656.2
2	NIEDERBR	65769.9	1938719	59717	43150053	65509.6	1362.11	62839.9	68179.4
3	HAGUENAU	69934.2	1635304	67266	61684374	70395.3	1281.02	67894.5	72906.1
4	SARRE-UN	63714.1	3111510	64704	241142599	58909.1	2301.35	53418.4	64399.7
5	SAVERNE	61219.3	2235967	63615	40468532	61347.9	1470.97	58464.8	64231.0
6	STRASBOU	72534.1	950069	78641	39590204	72677.2	953.22	70799.3	74565.1
7	MOLSHEIM	68356.3	1433369	72694	89306658	68437.4	1139.78	66105.4	70769.3
8	SCHIRMEC	63437.5	2436525	47787	92437865	62987.9	1539.01	59971.4	66004.3
9	SELESTAT	61764.6	1355503	63324	79340658	61799.0	1154.54	59536.1	64061.9
10	STE-MARI	63455.4	2380890	47025	107302666	63100.1	1526.17	60108.3	66091.4
11	COLMAR	69494.7	1770055	73872	50338419	69812.9	1307.66	67249.9	72375.9
12	NF-BRISS	63355.2	3672817	75507	97119251	63808.9	1881.22	60121.7	67496.1
13	SUESWILL	67950.1	2327083	73836	54594305	68192.9	1493.97	65264.7	71121.1
14	THANN-CE	68426.3	1637331	70133	29294290	67986.0	1245.26	65547.3	70428.7
15	MULHOUSE	70323.0	529532	72213	13036793	70395.8	713.35	68997.7	71794.0
16	ST-LOUIS	75350.7	742089	65540	51975448	75705.5	855.36	74029.0	77382.0
17	ALTKIRCH	66013.3	3353774	59055	163794602	65677.7	1813.40	62323.5	69432.0