Échantillonnage

Chapitre 13 : Régression, post-stratification et calage

Jaouad Madkour

30 octobre 2026

Plan du chapitre

Une meilleure droite **2**
Ce que le ratio suppose . . . 2
La droite de la population . . . 2
L'estimateur par la régression . . . 3
Sur la mini-population . . . 3

La précision de la régression **3**
La variance approchée . . . 3
Sur la population du cours . . . 5
Ratio ou régression . . . 5

L'échantillon fil rouge **6**
L'estimation . . . 6
La précision estimée . . . 6
Vingt mille échantillons . . . 7

La post-stratification **7**
Stratifier après le tirage . . . 7
Le fil rouge par milieu . . . 7
Post-stratifier, c'est repondérer . . . 8
Sa précision . . . 9

Le calage **10**
Le principe . . . 10
Le calage linéaire . . . 10
Les poids calés du fil rouge . . . 11
Choisir les variables de calage . . . 12

Le calage sur marges **12**
Quand on ne connaît que les marges . . . 12
Le tableau de l'échantillon . . . 12
Le premier tour . . . 13
Les itérations . . . 13
Les poids calés . . . 13

S'entraîner **14**
Exercice 1 : surface et chiffre d'affaires 14
Exercice 1 : solution 15
Exercice 2 : post-stratifier une enquête 15
Exercice 2 : solution 15
Exercice 3 : deux marges, quatre cases 16
Exercice 3 : solution 16

Ce qu'il faut retenir **16**
Le résumé du chapitre 16

Une meilleure droite

Ce que le ratio suppose

Rappel : l'estimateur par le ratio

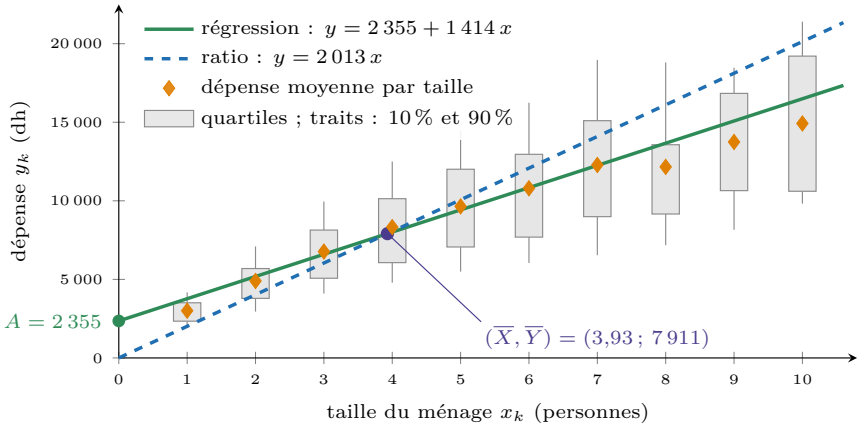
$$\widehat{\overline{Y}}_R = \overline{y}\,\frac{\overline{X}}{\overline{x}}$$

Il suppose la dépense **proportionnelle** à la taille, $y_k \approx R\,x_k$: une droite qui passe par l'origine. Avec la taille du ménage ($\overline{X} = 7\,859/2\,000 = 3{,}9295$, connu par le recensement), son effet de plan vaut 0,683 (chapitre 12).

Une personne de plus ne double pas la dépense

Les ménages d'une personne dépensent en moyenne 3 005 dh ; ceux de deux personnes 4 899 dh, et non 6 010. Une partie de la dépense (logement, équipement) ne dépend pas de la taille : la relation a une **ordonnée à l'origine**.

La droite de la population



Lire la figure

Les deux droites passent par le point moyen, avec des pentes de 2 013 dh par personne (ratio) et 1 414 (régression). Forcée par l'origine, celle du ratio est **trop basse** pour les petits

ménages, **trop haute** pour les grands.

L'estimateur par la régression

Corriger la moyenne par l'écart des tailles

$$\widehat{\overline{Y}}_{\text{reg}} = \overline{y} + b\,(\overline{X} - \overline{x}), \qquad b = \frac{s_{xy}}{s_x^2}$$

b est la pente des moindres carrés **dans l'échantillon**, avec $s_{xy} = \frac{1}{n-1}\sum_{k\in S}(x_k - \overline{x})(y_k - \overline{y})$. $\overline{X}$ est connu ; $\overline{x}$ et $\overline{y}$ sont mesurés.

La lecture

Si les ménages tirés sont trop petits ($\overline{x} < \overline{X}$), il leur manque $\overline{X} - \overline{x}$ personne en moyenne, et chaque personne vaut b dirhams : on ajoute $b\,(\overline{X} - \overline{x})$. Deux cas particuliers : $b = 0$ redonne $\overline{y}$; $b = \overline{y}/\overline{x}$ redonne le ratio. La régression **choisit la pente** sur les données.

Sur la mini-population

Tailles 2, 3, 3, 4, 6 ; trois ménages tirés

Population : $\overline{X} = 3{,}6$, $B = 2{,}5$, $A = -1$, $\rho^2 = 0{,}991$. Échantillon A, B, C : $\overline{y} = 17/3$, $\overline{x} = 8/3$, $b = 2{,}5$, et

$$\widehat{\overline{Y}}_{\text{reg}} = \frac{17}{3} + 2{,}5 \times \left(3{,}6 - \frac{8}{3}\right) = \frac{17}{3} + \frac{7}{3} = 8 \qquad \text{ratio} : \frac{17}{3} \times \frac{3{,}6}{8/3} = 7{,}65$$

Les dix échantillons possibles

Estimateur	Espérance	Variance	Erreur quadratique
Moyenne simple	8	1,933	1,933
Ratio	7,974	0,045	0,046
Régression	8	0,026	0,026

Sans biais ici (résidus de B et C opposés : $\mp 0{,}5$) ; en général, biais d'ordre $1/n$.

La précision de la régression

La variance approchée

Le résultat

$$V(\widehat{\overline{Y}}_{\text{reg}}) \approx \frac{1-f}{n}\,S_e^2 = \frac{1-f}{n}\,S^2\,(1-\rho^2)$$

S_e^2 est la variance des résidus $e_k = y_k - A - B\,x_k$ de la droite de la population, $B = S_{xy}/S_x^2$, $A = \overline{Y} - B\,\overline{X}$; ρ est la corrélation entre x et y.

La démonstration courte

Pour n grand, $b \approx B$, donc $\widehat{\overline{Y}}_{\text{reg}} \approx B\,\overline{X} + \overline{u}$, où $\overline{u}$ est la moyenne d'échantillon de $u_k = y_k - B\,x_k$. Sa variance est $(1-f)\,S_u^2/n$ (chapitre 3), et

$$S_u^2 = S^2 - 2B\,S_{xy} + B^2 S_x^2 = S^2 - \frac{S_{xy}^2}{S_x^2} = S^2\,(1-\rho^2)$$

Comme $u_k = e_k + A$, $S_u^2 = S_e^2$.

Démonstration : La régression, avec une pente fixée puis estimée

Une pente fixée. Soit b_0 un nombre fixé à l'avance, et $\widehat{\overline{Y}}(b_0) = \overline{y} + b_0\,(\overline{X} - \overline{x})$. Avec $u_k = y_k - b_0\,x_k$, on a $\widehat{\overline{Y}}(b_0) = b_0\,\overline{X} + \overline{u}$: un sondage aléatoire simple de la variable u. Il est donc **sans biais**, de variance

$$V\big(\widehat{\overline{Y}}(b_0)\big) = \frac{1-f}{n}\,S_u^2, \qquad S_u^2 = S^2 - 2b_0\,S_{xy} + b_0^2\,S_x^2.$$

La meilleure pente. En complétant le carré, avec $B = S_{xy}/S_x^2$,

$$S_u^2 = S^2 - \frac{S_{xy}^2}{S_x^2} + S_x^2\,(b_0 - B)^2 = S^2\,(1-\rho^2) + S_x^2\,(b_0 - B)^2,$$

car $S_{xy}^2/S_x^2 = \rho^2\,S^2$. Cette variance est minimale pour $b_0 = B$, où elle vaut $S^2\,(1-\rho^2)$. Pour $b_0 = 0$, on retrouve $\overline{y}$; pour $b_0 = R$, le ratio.

La pente estimée. L'estimateur du cours utilise la pente b de l'échantillon. Posons $u_k = y_k - B\,x_k$ et $\overline{U} = \overline{Y} - B\,\overline{X}$. On vérifie, en développant,

$$\widehat{\overline{Y}}_{\text{reg}} - \overline{Y} = (\overline{u} - \overline{U}) - (b - B)\,(\overline{x} - \overline{X}).$$

Le premier terme est l'erreur d'un sondage simple, d'ordre $1/\sqrt{n}$. Le second est le produit de deux erreurs d'ordre $1/\sqrt{n}$ chacune, donc d'ordre $1/n$: négligeable. Au premier ordre, la variance est $\frac{1-f}{n}\,S_u^2 = \frac{1-f}{n}\,S^2\,(1-\rho^2)$.

Ce que cela dit. Sur la mini-population, $B = 2{,}5$, $A = -1$, $\rho^2 = 0{,}991$; l'énumération des échantillons de deux et trois ménages retrouve, pour cinq pentes fixées, l'absence de biais et la variance de la formule, et pour la pente estimée l'identité ci-dessus sur chacun des dix échantillons de trois ménages ; l'estimateur y a pour espérance 8 et pour variance 0,026, contre 1,933 pour $\overline{y}$ et 0,045 pour le ratio. Sur la population du cours, $S^2\,(1-\rho^2)$ donne $S_e = 3\,120{,}0$, une variance de $0{,}0045 \times 3\,120{,}0^2 = 43\,805{,}5$, une erreur-type de 209,3 dh, soit 21,7 % de moins que $\overline{y}$, et une taille effective de $200/0{,}613 \approx 326$.

■

Sur la population du cours

Dépense et taille, 200 ménages

$B = 4\,342{,}39/3{,}07107 = 1\,413{,}97$ dh par personne ; $A = 2\,355{,}25$ dh ; $\rho = 0{,}622$.

	Moyenne simple	Ratio	Régression
Écart-type des résidus	3 984,3	3 292,1	3 120,0
Variance	71 435,5	48 770,3	43 805,5
Erreur-type	267,3	220,8	209,3
Effet de plan	1	0,683	0,613
Taille effective	200	293	326

Deux cents ménages qui en valent 326

deff $= 1 - \rho^2 = 0{,}613$: interroger 200 ménages et utiliser leur taille donne la précision de 326 ménages tirés sans information auxiliaire. L'erreur-type baisse de 21,7 %, sans un questionnaire de plus.

Ratio ou régression

La régression ne perd pas

Pour n grand,

$$V(\widehat{\overline{Y}}_R) - V(\widehat{\overline{Y}}_{\text{reg}}) \approx \frac{1-f}{n}\,(R-B)^2\,S_x^2 \;\geq\; 0$$

Les deux coïncident si $R = B$, c'est-à-dire si la droite des moindres carrés passe par l'origine ($A = 0$).

Dépense et taille

$R - B = 2\,013{,}34 - 1\,413{,}97 = 599{,}38$, et

$$0{,}0045 \times 599{,}38^2 \times 3{,}0711 \approx 4\,965 \approx 48\,770{,}3 - 43\,805{,}5$$

En effets de plan, l'écart vaut $(\rho - k)^2 = (0{,}6219 - 0{,}8855)^2 = 0{,}0695$, avec $k = CV_x/CV_y$ (chapitre 12) : c'est $0{,}6827 - 0{,}6132$.

Démonstration : La régression ne perd jamais devant le ratio

La différence exacte. Pour une pente fixée b_0, la variance de l'estimateur est $\frac{1-f}{n}[S^2(1-\rho^2) + S_x^2\,(b_0 - B)^2]$ (démonstration précédente). Le ratio correspond à $b_0 = R$, la régression à $b_0 = B$:

$$V(\widehat{\overline{Y}}_R) - V(\widehat{\overline{Y}}_{\text{reg}}) \approx \frac{1-f}{n}\,S_x^2\,(R-B)^2 \;\geq\; 0.$$

Quand l'égalité ? Si et seulement si $R = B$. Or $A = \overline{Y} - B\,\overline{X} = \overline{X}\,(R - B)$: l'égalité équivaut à $A = 0$, une droite des moindres carrés qui passe par l'origine.

En effets de plan. On divise par $\frac{1-f}{n}\,S^2$: $\dfrac{S_x^2\,(R-B)^2}{S^2}$. Avec $R\,S_x/S = k$ et $B\,S_x/S = \rho$, c'est

$(k-\rho)^2$, l'écart de la démonstration du chapitre 12.

Ce que cela dit. Sur la population du cours, $R - B = 2\,013{,}34 - 1\,413{,}97 = 599{,}38$ et $0{,}0045 \times 599{,}38^2 \times 3{,}0711 \approx 4\,965 = 48\,770{,}3 - 43\,805{,}5$. En effets de plan, $(\rho - k)^2 = 0{,}0695$ et $0{,}6827 - 0{,}6132 = 0{,}0695$. Sur la mini-population, l'énumération confirme $V_R - V_{\text{reg}} = \frac{1-f}{n} S_x^2 (R-B)^2$ pour $n = 2$ et $n = 3$, pour la pente fixée.

■

L'échantillon fil rouge

L'estimation

Les 200 ménages du chapitre 3

$\overline{x} = 3{,}805$ personnes, $s_x^2 = 3{,}1728$, $s_{xy} = 4\,630{,}2$, d'où $b = 1\,459{,}34$ dh par personne.

$$\widehat{\overline{Y}}_{\text{reg}} = 7\,843{,}05 + 1\,459{,}34 \times (3{,}9295 - 3{,}805) = 7\,843{,}05 + 181{,}69 = 8\,024{,}74 \text{ dh}$$

Trois estimations, une seule vérité

Estimateur	Estimation	Erreur
Moyenne simple $\overline{y}$	7 843,05	−68,4
Ratio	8 099,68	+188,2
Régression	8 024,74	+113,3

Ménages tirés trop petits : la régression relève l'estimation **moins** que le ratio (pente 1 459 au lieu de 2 061), mais encore trop. Ici, $\overline{y}$ était, par chance, le plus proche.

La précision estimée

Estimer la variance sur l'échantillon

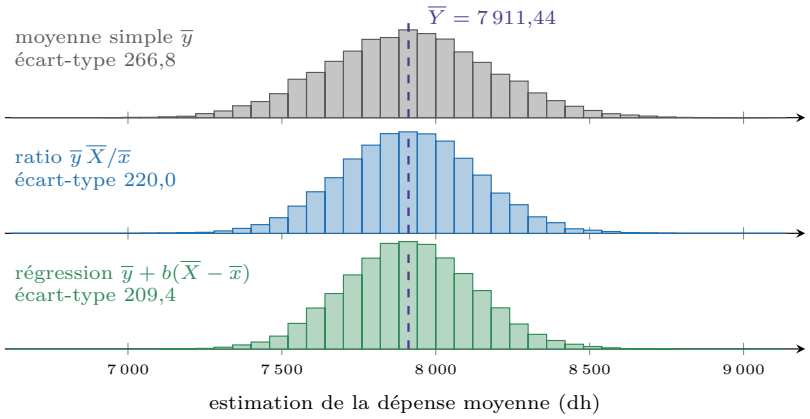
$$\widehat{V}(\widehat{\overline{Y}}_{\text{reg}}) = \frac{1-f}{n} s_e^2, \qquad s_e^2 = \frac{1}{n-2} \sum_{k \in S} (y_k - a - b\,x_k)^2, \quad a = \overline{y} - b\,\overline{x}$$

On divise par $n - 2$: deux coefficients, a et b, ont été estimés sur l'échantillon.

L'échantillon fil rouge

$s_e = 2\,818{,}7$ dh, d'où une erreur-type estimée de $\sqrt{0{,}0045} \times 2\,818{,}7 = 189{,}1$ dh. Intervalle à 95 % : $8\,024{,}74 \pm 370{,}6$, soit $[7\,654{,}1\,;8\,395{,}3]$, qui contient $\overline{Y} = 7\,911{,}44$. Sa largeur, 741 dh, est celle de l'intervalle de $\overline{y}$ (1 007 dh, chapitre 3) diminuée d'un quart.

Vingt mille échantillons



Les trois estimateurs, 20 000 tirages de 200 ménages

Trois lois centrées sur $\overline{Y}$ (biais de 2 dh pour le ratio et la régression), d'écarts-types 266,8 ; 220,0 ; 209,4 dh, comme l'annonçaient les formules. La régression tombe à moins de 300 dh de $\overline{Y}$ dans 84,9 % des tirages, contre 73,7 % pour $\overline{y}$.

La post-stratification

Stratifier après le tirage

La post-stratification

On tire un sondage aléatoire simple. Après la collecte, on classe les unités tirées en groupes h dont les effectifs N_h sont connus (recensement), et l'on estime

$$\widehat{\overline{Y}}_{\text{post}} = \sum_h W_h\,\overline{y}_h, \qquad W_h = \frac{N_h}{N}$$

Une stratification dont les effectifs sont aléatoires

La formule est celle du sondage stratifié (chapitre 6), mais les n_h ne sont pas fixés à l'avance : ils dépendent du tirage. On post-stratifie quand on ne peut pas stratifier : la variable ne figure pas dans la base de sondage, ou on ne la connaît qu'à l'entretien.

Le fil rouge par milieu

48 ruraux au lieu de 56

Milieu	N_h	W_h	n_h	$n\,W_h$	$\overline{y}_h$
Urbain	1 438	0,719	152	143,8	8 135,59
Rural	562	0,281	48	56,2	6 916,67

Estimation : $\widehat{\overline{Y}}_{\text{post}} = 0{,}719 \times 8\,135{,}59 + 0{,}281 \times 6\,916{,}67 = 7\,793{,}07$ dh.

Rendre aux ruraux leur part

24 % de ruraux dans l'échantillon, 28,1 % dans la population : $\overline{y} = 7\,843{,}05$ pèse trop les urbains. La post-stratification corrige ce déséquilibre, **et seulement lui**. Ici, elle éloigne de $\overline{Y}$ (erreur $-118{,}4$ au lieu de $-68{,}4$) : les urbains tirés dépensent moins que l'ensemble des urbains (8 136 contre 8 430 dh).

Post-stratifier, c'est repondérer

Les poids de la post-stratification

Chaque unité tirée du groupe h reçoit le poids $w_k = N_h/n_h$ au lieu de $d_k = N/n$. Alors

$$\widehat{\overline{Y}}_{\text{post}} = \frac{1}{N}\sum_{k\in S} w_k\, y_k \qquad \text{et} \qquad \sum_{k\in S_h} w_k = N_h$$

Les poids reproduisent **exactement** les effectifs connus.

Sur le fil rouge

Urbains : $1\,438/152 = 9{,}461$; ruraux : $562/48 = 11{,}708$, au lieu de 10 pour tous. La somme des poids vaut $152 \times \frac{1\,438}{152} + 48 \times \frac{562}{48} = 2\,000 = N$. C'est déjà un **calage** : des poids retouchés pour retrouver des totaux connus.

Démonstration : La post-stratification, des poids qui reproduisent les effectifs

Une moyenne repondérée. Chaque unité tirée du groupe h reçoit le poids $w_k = N_h/n_h$. Alors

$$\sum_{k\in S_h} w_k = n_h\,\frac{N_h}{n_h} = N_h, \qquad \frac{1}{N}\sum_{k\in S} w_k\, y_k = \frac{1}{N}\sum_h \frac{N_h}{n_h}\sum_{k\in S_h} y_k = \sum_h W_h\,\overline{y}_h.$$

Les poids reproduisent **exactement** les effectifs N_h, et la moyenne repondérée est l'estimateur $\sum_h W_h\,\overline{y}_h$.

Pas de biais, sachant les effectifs. Si l'on connaît les effectifs $n_1, \dots, n_H$ de l'échantillon, les n_h unités tirées dans le groupe h forment un sondage aléatoire simple de n_h unités parmi les N_h du groupe : tous les échantillons de même composition sont également probables. Donc $E(\overline{y}_h \mid n_1, \dots, n_H) = \overline{Y}_h$ pour chaque groupe représenté, et

$$E\Big(\sum_h W_h\,\overline{y}_h \,\Big|\, n_1, \dots, n_H\Big) = \sum_h W_h\,\overline{Y}_h = \overline{Y}.$$

Il suffit que chaque groupe soit représenté ($n_h \geq 1$) : sinon la moyenne $\overline{y}_h$ n'existe pas. Avec 200 ménages, la probabilité qu'aucun rural ne soit tiré est de l'ordre de 3×10^{-31}.

Ce que cela dit. Sur le fil rouge, 9,461 pour les 152 urbains et 11,708 pour les 48 ruraux : $\sum_S w_k = 2\,000$ et $0{,}719 \times 8\,135{,}59 + 0{,}281 \times 6\,916{,}67 = 7\,793{,}07$ dh. Sur la mini-population, avec les groupes ABC et DE et des échantillons de trois ménages, l'énumération des neuf échantillons qui représentent les deux groupes donne $E(\overline{y}_{\text{post}}) = 8 = \overline{Y}$.

■

Sa précision

Presque la stratification proportionnelle

$$V(\widehat{\overline{Y}}_{\text{post}}) \approx \underbrace{\frac{1-f}{n}\sum_h W_h\, S_h^2}_{\text{stratifié proportionnel}} + \frac{1}{n^2}\sum_h (1-W_h)\, S_h^2$$

Le second terme, d'ordre $1/n^2$, est le prix des effectifs n_h aléatoires.

Par milieu, 200 ménages

$V \approx 68\,367{,}8 + 305{,}9 = 68\,673{,}7$: effet de plan 0,961, contre 0,957 pour la stratification proportionnelle (chapitre 6). Sur le fil rouge, $\widehat{V} = \sum_h W_h^2\,(1 - n_h/N_h)\, s_h^2/n_h$ donne une erreur-type de 253,1 dh. Le milieu n'explique que 4,3 % de la variance de la dépense : il fait bien moins que la taille du ménage (effet de plan 0,613).

Démonstration : La variance de la post-stratification

Sachant les effectifs. Les groupes sont des sondages simples indépendants, de tailles n_h fixées par le conditionnement : $V(\overline{y}_{\text{post}} \mid n_h) = \sum_h W_h^2\,(\frac{1}{n_h} - \frac{1}{N_h})S_h^2$. L'espérance conditionnelle $\overline{Y}$ est constante, donc la loi de la variance totale donne $V(\overline{y}_{\text{post}}) = E[V(\overline{y}_{\text{post}} \mid n_h)]$: il reste à moyenner $1/n_h$.

L'espérance de $1/n_h$. L'effectif n_h est hypergéométrique : $E(n_h) = n\,W_h$ et $V(n_h) = n\,W_h\,(1 - W_h)\,\frac{N-n}{N-1} \approx n\,W_h\,(1 - W_h)\,(1 - f)$. La fonction $1/x$ étant convexe, le développement de Taylor à l'ordre deux donne

$$E\Big(\frac{1}{n_h}\Big) \approx \frac{1}{E(n_h)} + \frac{V(n_h)}{E(n_h)^3} = \frac{1}{n\,W_h} + \frac{(1-W_h)\,(1-f)}{n^2\,W_h^2}.$$

Le résultat. Avec $N_h = N\,W_h$, $\frac{1}{n\,W_h} - \frac{1}{N_h} = \frac{1-f}{n\,W_h}$, d'où

$$V(\overline{y}_{\text{post}}) \approx \frac{1-f}{n}\sum_h W_h\, S_h^2 + \frac{1-f}{n^2}\sum_h (1-W_h)\, S_h^2.$$

Le premier terme est la variance du plan stratifié proportionnel. Le second, d'ordre $1/n^2$, est le prix de l'aléa des n_h ; le cours le simplifie en remplaçant $1 - f$ par 1.

Ce que cela dit. Pour le milieu et $n = 200$, le premier terme vaut 68 367,8 et le second 305,9 (avec $1 - f$: 275,3), soit 68 673,7 et un effet de plan de 0,961 ; avec $1 - f$, 68 643,1. Le calcul exact avec la loi hypergéométrique des n_h donne 68 649,0 : la simplification du cours change la variance de 0,04 %. Sur la mini-population, l'énumération des neuf échantillons confirme $V(\overline{y}_{\text{post}}) = E\big[\sum_h W_h^2\,(\frac{1}{n_h} - \frac{1}{N_h})S_h^2\big]$. L'estimateur $\widehat{V} = \sum_h W_h^2\,(1 - n_h/N_h)\, s_h^2/n_h$ donne, sur le fil rouge, une erreur-type de 253,1 dh.

■

Le calage

Le principe

Caler les poids

On cherche des poids w_k **aussi proches que possible** des poids de sondage d_k, qui reproduisent des totaux connus :

$$\sum_{k\in S} w_k\, x_k = X \quad \text{pour chaque variable auxiliaire } x$$

puis l'on estime $\widehat{Y}_{\text{cal}} = \sum_{k\in S} w_k\, y_k$.

Un seul jeu de poids pour toute l'enquête

Les mêmes poids servent pour la dépense, l'accès à internet ou toute autre variable. Toutes les estimations retrouvent les chiffres connus (ménages, personnes, ruraux) : l'enquête devient **cohérente** avec le recensement. La post-stratification est le calage sur les effectifs des groupes.

Le calage linéaire

Des poids linéaires en x

On impose $\sum_S w_k = N$ et $\sum_S w_k\, x_k = X$ à des poids $w_k = d_k\,(1 + \lambda_1 + \lambda_2\, x_k)$. Pour un sondage aléatoire simple :

$$w_k = \frac{N}{n}[1 + c\,(x_k - \overline{x})], \qquad c = \frac{\overline{X} - \overline{x}}{\hat{\sigma}_x^2}$$

où $\hat{\sigma}_x^2 = \frac{1}{n}\sum_S (x_k - \overline{x})^2$ est la variance de x dans l'échantillon, divisée par n.

La démonstration courte

$\sum_S (x_k - \overline{x}) = 0$ donne $\sum_S w_k = N$ et $\sum_S w_k\, x_k = N\,\overline{x} + Nc\,\hat{\sigma}_x^2 = N\,\overline{X}$. Avec $\hat{\sigma}_{xy} = \frac{1}{n}\sum_S (x_k - \overline{x})\, y_k$, la moyenne calée est **l'estimateur par la régression** :

$$\frac{1}{N}\sum_{k\in S} w_k\, y_k = \overline{y} + c\,\hat{\sigma}_{xy} = \overline{y} + \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}\,(\overline{X} - \overline{x}) = \overline{y} + b\,(\overline{X} - \overline{x})$$

Démonstration : Le calage linéaire, des poids qui redonnent la régression

Pourquoi des poids linéaires en x. On cherche les poids w_k les plus proches des poids de sondage $d_k = d = N/n$, au sens de la distance du khi-deux $\sum_{k\in S}(w_k - d)^2/(2d)$, avec les deux contraintes $\sum_S w_k = N$ et $\sum_S w_k\, x_k = X$. Avec des multiplicateurs de Lagrange λ_1 et λ_2, l'annulation de la dérivée en w_k donne

$$\frac{w_k - d}{d} = \lambda_1 + \lambda_2\, x_k, \qquad \text{soit} \qquad w_k = d\,(1 + \lambda_1 + \lambda_2\, x_k).$$

Les contraintes. Comme $n\,d = N$, la contrainte $\sum_S w_k = N$ donne $\lambda_1 + \lambda_2\,\overline{x} = 0$, donc $w_k =$

$d\,[1 + c\,(x_k - \overline{x})]$ avec $c = \lambda_2$. Avec $\sum_S (x_k - \overline{x})\,x_k = n\,\hat{\sigma}_x^2$,

$$\sum_S w_k\,x_k = N\,\overline{x} + N\,c\,\hat{\sigma}_x^2 = N\,\overline{X} \quad \Longrightarrow \quad c = \frac{\overline{X} - \overline{x}}{\hat{\sigma}_x^2}.$$

C'est la régression. Avec $\hat{\sigma}_{xy} = \frac{1}{n}\sum_S (x_k - \overline{x})\,y_k$, on a $\frac{1}{N}\sum_S w_k\,y_k = \overline{y} + c\,\hat{\sigma}_{xy}$. Les diviseurs n de $\hat{\sigma}_{xy}$ et de $\hat{\sigma}_x^2$ se simplifient avec ceux de s_{xy} et de s_x^2 : $c\,\hat{\sigma}_{xy} = (\overline{X} - \overline{x})\,s_{xy}/s_x^2 = b\,(\overline{X} - \overline{x})$.

Les poids sont les plus proches. Soit w'_k d'autres poids qui reproduisent les deux totaux, et $\delta_k = w'_k - w_k$: $\sum \delta_k = 0$ et $\sum \delta_k\,x_k = 0$, donc $\sum \delta_k\,(w_k - d)/d = c\sum \delta_k\,(x_k - \overline{x}) = 0$. Il reste

$$\sum_S \frac{(w'_k - d)^2}{d} = \sum_S \frac{(w_k - d)^2}{d} + \sum_S \frac{\delta_k^2}{d} \geq \sum_S \frac{(w_k - d)^2}{d},$$

avec égalité si et seulement si $w' = w$.

Ce que cela dit. Sur le fil rouge, $\overline{x} = 3{,}805$ et $\hat{\sigma}_x^2 = 3{,}1570$: $c = 0{,}1245/3{,}1570 = 0{,}0394$; les poids vont de 8,89 à 12,44, leur somme vaut 2 000, $\sum w_k\,x_k = 7\,859$, et $\frac{1}{N}\sum_S w_k\,y_k = 8\,024{,}74$, exactement l'estimation par la régression. Sur 200 échantillons tirés au hasard dans la population, les contraintes sont reproduites et la moyenne calée égale l'estimateur par la régression ; sur 50 jeux de poids admissibles tirés au hasard, la distance du khi-deux ne bat jamais celle du calage linéaire.

■

Les poids calés du fil rouge

Il manque 249 personnes

Avec $d_k = 10$, les 200 ménages « représentent » $10 \times 761 = 7\,610$ personnes, contre $X = 7\,859$. Ici $c = 0{,}1245/3{,}1570 = 0{,}0394$:

Taille x_k	1	3	4	6	10
Poids	8,89	9,68	10,08	10,87	12,44

$\sum_S w_k = 2\,000$, $\sum_S w_k\,x_k = 7\,859$, et $\frac{1}{N}\sum_S w_k\,y_k = 8\,024{,}74$: l'estimation par la régression.

Des poids négatifs

Si l'échantillon est très déséquilibré, ou si l'on cale sur beaucoup de variables, des poids linéaires peuvent devenir très grands, voire **négatifs** : un ménage compterait pour moins que rien. On borne alors les poids, ou l'on utilise un calage multiplicatif, qui les garde positifs. Ici, ils restent entre 8,89 et 12,44.

Choisir les variables de calage

Ce que rapporte chaque information

Sans la taille	$1 - R^2$	Avec la taille	$1 - R^2$
Milieu	0,957	Taille seule	0,613
Province	0,953	Taille et milieu	0,490
Milieu et province	0,930	Les trois	0,464

R^2 : part de la variance de la dépense expliquée par ces variables dans la population. L'effet de plan du calage vaut environ $1 - R^2$ (simulé pour les trois variables : 0,470).

Une bonne variable de calage

Elle est **liée** à la variable étudiée ; son total est connu **exactement** (recensement, fichier administratif) ; elle est mesurée de la même façon dans l'enquête et dans la source.

Le calage sur marges

Quand on ne connaît que les marges

Le calage sur marges

On connaît les effectifs par milieu (1 438 urbains, 562 ruraux) et par province (820, 560, 380, 240), mais pas le tableau croisé milieu × province. On cherche des poids qui reproduisent **les deux marges à la fois**.

L'ajustement proportionnel itératif

1. Partir des poids de sondage d_k.
2. Multiplier les poids de chaque milieu par un même facteur, pour retrouver sa marge.
3. Faire de même pour chaque province.
4. Recommencer 2 et 3 : chaque étape dérange un peu l'autre marge, de moins en moins.

Les poids, multipliés à chaque étape, restent **positifs**.

Le tableau de l'échantillon

Les 200 ménages du fil rouge, avec le poids 10

	Tanger-Assilah	Tétouan	Larache	Al Hoceïma	Total
Urbains tirés	70	49	20	13	152
Ruraux tirés	13	11	9	15	48
Estimé (×10)	830	600	290	280	2 000
Connu	820	560	380	240	2 000

Milieu : 1 520 urbains et 480 ruraux estimés, pour 1 438 et 562.

Deux déséquilibres à corriger

Larache est sous-représentée (76 % de sa marge), Al Hoceïma sur-représentée (117 %), les ruraux sous-représentés (85 %).

Le premier tour

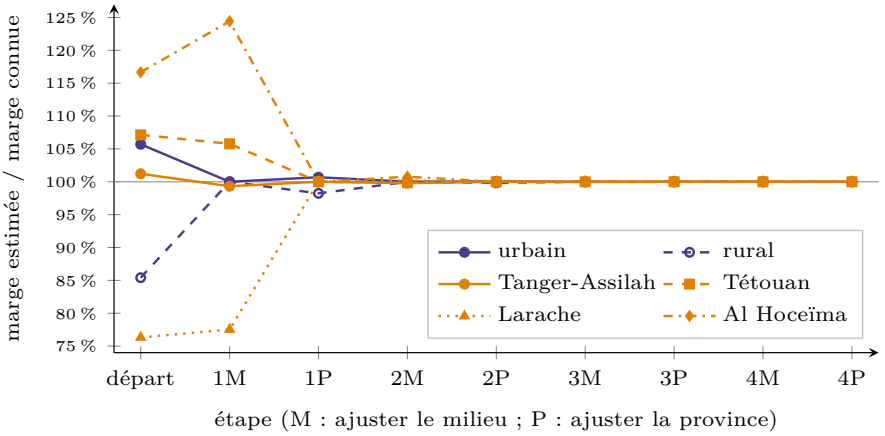
Étape 1 : le milieu

Urbains : poids $\times 1\,438/1\,520 = 0{,}9461$; ruraux : $\times 562/480 = 1{,}1708$. Le milieu est juste ; les provinces deviennent 814,45 ; 592,36 ; 294,59 ; 298,61.

Étape 2 : la province

Facteurs $820/814{,}45 = 1{,}0068$; $560/592{,}36 = 0{,}9454$; $380/294{,}59 = 1{,}2899$; $240/298{,}61 = 0{,}8037$. Les provinces sont justes ; le milieu devient 1 447,9 urbains et 552,1 ruraux, à 10 ménages près.

Les itérations



Une convergence rapide

Au départ, les marges estimées s'écartent jusqu'à 24 % des marges connues. Après un tour, plus aucune ne s'en écarte de 10 ménages ; après deux tours, d'un ménage (0,86) ; après quatre, d'un centième de ménage.

Les poids calés

Après quatre tours

Poids w_k	Tanger-Assilah	Tétouan	Larache	Al Hoceïma
Urbain	9,48	8,89	12,09	7,48
Rural	12,05	11,30	15,37	9,51

Estimation : $\frac{1}{N}\sum_S w_k\, y_k = 7\,852{,}13$ dh, une erreur de $-59{,}3$ dh.

Un facteur par milieu, un facteur par province

Chaque poids est le produit $10\times$ (facteur du milieu) $\times$ (facteur de la province) : dans chaque province, un rural pèse 1,271 fois un urbain. Sur 20 000 tirages, l'effet de plan vaut 0,943 : mieux que le milieu seul (0,957) ou la province seule (0,962), loin de la taille (0,613).

Démonstration : L'ajustement proportionnel itératif

Une étape rend une marge exacte. On multiplie tous les poids des unités du groupe i de la première variable par m_i/t_i, où m_i est la marge connue et t_i la marge actuelle : la nouvelle marge vaut $t_i \times m_i/t_i = m_i$. L'étape suivante fait de même pour la seconde variable.

La forme produit. Chaque étape multiplie les poids par un facteur qui dépend seulement du groupe de la variable ajustée. Après quelques tours, le poids d'une unité du milieu i et de la province j est donc

$$w_k = d \times a_i \times b_j,$$

où a_i est le produit des facteurs reçus par le milieu i et b_j celui des facteurs de la province j. Trois conséquences : les poids restent **strictement positifs** ; dans une même province, le rapport entre les poids d'un rural et d'un urbain est $a_{\text{rural}}/a_{\text{urbain}}$, **le même** dans toutes les provinces ; le calage sur marges n'invente pas d'interaction entre le milieu et la province.

La convergence. Chaque étape dérange la marge de l'autre variable, de moins en moins. Pour des tableaux de poids strictement positifs, les facteurs convergent vers un tableau qui reproduit les deux marges à la fois : c'est un théorème d'Ireland et Kullback (1968), que l'on admet ici. L'écart maximal aux marges permet de savoir quand s'arrêter.

Ce que cela dit. Sur le fil rouge, les facteurs du premier tour sont $1\,438/1\,520 = 0{,}9461$ et $562/480 = 1{,}1708$ pour le milieu, puis 1,0068, 0,9454, 1,2899 et 0,8037 pour la province ; l'écart maximal aux marges vaut 9,92 ménages après un tour, 0,86 après deux, 0,07 après trois et 0,006 après quatre. Un rural y pèse 1,271 fois un urbain dans chacune des quatre provinces, et l'estimation est 7 852,13 dh. Dans l'exercice 3, le premier tour donne le tableau $\left(\begin{smallmatrix} 293{,}3 & 188{,}6 \\ 106{,}7 & 411{,}4 \end{smallmatrix}\right)$, dont le milieu reste déréglé (481,9 et 518,1) ; après cinq tours, $\left(\begin{smallmatrix} 316{,}9 & 233{,}1 \\ 83{,}1 & 366{,}9 \end{smallmatrix}\right)$.

■

S'entraîner

Exercice 1 : surface et chiffre d'affaires

Quatre commerces d'un souk

On tire 4 commerces parmi les $N = 120$ d'un souk. On mesure leur surface x (dizaines de m^2) et leur chiffre d'affaires mensuel y (milliers de dh) :

Commerce	1	2	3	4
Surface x_k	2	4	5	9
Chiffre d'affaires y_k	30	44	52	74

La surface moyenne des 120 commerces est connue : $\overline{X} = 6$.

1. Calculer $\overline{x}$, $\overline{y}$ et la pente b.
2. Estimer le chiffre d'affaires moyen et total par la régression, puis par le ratio.
3. Quelle estimation préférer ?

Exercice 1 : solution

La pente

1. $\overline{x} = 5$, $\overline{y} = 50$. Écarts : -3, -1, 0, 4 et -20, -6, 2, 24. Somme des produits $60+6+0+96 = 162$; somme des carrés $9+1+0+16 = 26$; $b = 162/26 = 6{,}23$.

Deux estimations

2. Régression : $50 + 6{,}23 \times (6-5) = 56{,}23$, total $120 \times 56{,}23 = 6\,747{,}7$ milliers de dh. Ratio : $50 \times 6/5 = 60$, total 7 200.
3. L'ordonnée à l'origine $a = 50 - 6{,}23 \times 5 = 18{,}85$ est loin de 0 : un commerce minuscule fait encore des ventes. La droite ne passe pas par l'origine, et la régression convient mieux. Avec 4 commerces seulement, les deux estimations restent fragiles.

Exercice 2 : post-stratifier une enquête

Cent ménages, deux milieux

Une commune compte $N = 1\,000$ ménages : 600 urbains et 400 ruraux. Un sondage aléatoire simple de 100 ménages a donné 70 urbains (dépense moyenne 9 milliers de dh, écart-type 3) et 30 ruraux (moyenne 5, écart-type 2).

1. Calculer la moyenne simple $\overline{y}$.
2. Calculer l'estimation post-stratifiée. Pourquoi diffère-t-elle ?
3. Donner les poids de la post-stratification et vérifier leur somme.
4. Estimer la variance de l'estimation post-stratifiée.

Exercice 2 : solution

Les estimations

1. $\overline{y} = (70 \times 9 + 30 \times 5)/100 = 7{,}8$.
2. $0{,}6 \times 9 + 0{,}4 \times 5 = 7{,}4$. L'échantillon compte 70 % d'urbains au lieu de 60 % : $\overline{y}$ surpondère les ménages qui dépensent le plus.
3. Urbains : $600/70 = 8{,}57$; ruraux : $400/30 = 13{,}33$, au lieu de 10. Somme : $70 \times \frac{600}{70} + 30 \times \frac{400}{30} = 1\,000$.

La variance

4. $0{,}6^2 \times (1 - \frac{70}{600}) \times \frac{9}{70} + 0{,}4^2 \times (1 - \frac{30}{400}) \times \frac{4}{30} = 0{,}0409 + 0{,}0197 = 0{,}0606$: erreur-type 0,246 millier de dh.

Exercice 3 : deux marges, quatre cases

Cent entreprises tirées sur mille

Une enquête tire 100 entreprises parmi 1 000 :

	Commerce	Agriculture
Urbain	30	10
Rural	20	40

On sait qu'il y a 550 entreprises urbaines et 450 rurales, 400 commerces et 600 exploitations agricoles.

1. Quelles marges donnent les poids $d_k = 10$?
2. Faire un tour d'ajustement proportionnel : milieu, puis secteur.
3. Le tableau estimé après ce tour reproduit-il les deux marges ?

Exercice 3 : solution

Le premier tour

1. Urbain 400, rural 600 ; commerce 500, agriculture 500.
2. Milieu : urbains $\times\, 550/400 = 1{,}375$ (poids 13,75), ruraux $\times\, 450/600 = 0{,}75$ (poids 7,5). Secteurs : commerce $412{,}5 + 150 = 562{,}5$, agriculture 437,5. Facteurs $400/562{,}5 = 0{,}711$ et $600/437{,}5 = 1{,}371$.

Le tableau

3. $\begin{pmatrix} 293{,}3 & 188{,}6 \\ 106{,}7 & 411{,}4 \end{pmatrix}$: les secteurs sont justes, mais le milieu est déréglé (481,9 et 518,1). Il faut recommencer : après cinq tours, $\begin{pmatrix} 316{,}9 & 233{,}1 \\ 83{,}1 & 366{,}9 \end{pmatrix}$, qui reproduit les deux marges à un dixième près.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. L'estimateur par la **régression** : $\overline{y} + b\,(\overline{X} - \overline{x})$, b pente de l'échantillon.
2. Sa variance est $\frac{1-f}{n}\, S^2\,(1-\rho^2)$: effet de plan 0,613 pour la dépense et la taille.
3. Il fait au moins aussi bien que le ratio (n grand), et autant si $A = 0$.
4. **Post-stratification** : $\sum_h W_h\, \overline{y}_h$, poids N_h/n_h ; presque la précision d'un stratifié proportionnel.
5. **Calage** : des poids proches des d_k qui reproduisent des totaux connus.
6. Calage linéaire = régression ; calage sur marges : ajustement proportionnel itératif.
7. Effet de plan $\approx 1 - R^2$: caler sur des variables liées à y, connues exactement.

La suite

Le chapitre 14 traite des ménages tirés qui ne répondent pas. Les poids de ce chapitre y servent à nouveau : on **repondère** les répondants pour corriger la non-réponse.