MÉTHODES DE REDRESSEMENT OU DE PONDÉRATION

- Principe:

utiliser *a posteriori* une information supplémentaire corrélée avec la variable à étudier.

- Information:

variables de contrôle dont on connaît des caractéristiques globales, ou des caractéristiques par classes, ou des valeurs pour chaque unité.

MÉTHODES DE REDRESSEMENT OU DE PONDÉRATION

- Estimation par le quotient ou redressement par variable quantitative

Exemple:

- Échantillon de 80 hypermarchés - On veut estimer le CA moyen $\overline{Y}$
- On a $\overline{y} = 110,2\,k€$
- On sait que le nombre moyen $\overline{X}$ de caisses dans la population des hypermarchés est 28.
- Dans l'échantillon $\overline{x} = 28.8$

$$\hat{\overline{Y}} = 110.2 \times \frac{28}{28.8} = 107.1$$

Estimation par le quotient

Formule générale: $\bar{y}_q = \bar{y}\frac{\bar{X}}{\bar{x}}$

Remarque: en général estimation biaisée, mais biais négligeable si n>1000.

Hypothèse de proportionnalité (règle de 3)

Estimation par le quotient

- Calcul du biais:

$$\overline{y_q} = \overline{X}\frac{\overline{y}}{\overline{x}} = \overline{X}\frac{\overline{y}-\overline{Y}+\overline{Y}}{\overline{x}-\overline{X}+\overline{X}} = \overline{Y}\frac{1+\frac{\overline{y}-\overline{Y}}{\overline{Y}}}{\underbrace{1+\frac{\overline{x}-\overline{X}}{\overline{X}}}_{\varepsilon}}$$

Développement limité:

$$\overline{y_q} \simeq \overline{Y}\left(1+\frac{\overline{y}-\overline{Y}}{\overline{Y}}\right)\left[1-\frac{\overline{x}-\overline{X}}{\overline{X}}+\left(\frac{\overline{x}-\overline{X}}{\overline{X}}\right)^2\right] \simeq \overline{Y}\left[1+\frac{\overline{y}-\overline{Y}}{\overline{Y}}-\frac{\overline{x}-\overline{X}}{\overline{X}}\times\frac{\overline{y}-\overline{Y}}{\overline{Y}}-\frac{\overline{x}-\overline{X}}{\overline{X}}+\left(\frac{\overline{x}-\overline{X}}{\overline{X}}\right)^2\right]$$

$$E\left(\bar{y}_q\right) \simeq \bar{Y}\left[1 - \frac{\mathrm{cov}\left(\bar{x};\bar{y}\right)}{\bar{X}\bar{Y}} + \frac{V\left(\bar{x}\right)}{\bar{X}^2}\right]$$

Si probabilité égale et sans remise:

$$E\left(\bar{y}_q\right) = \bar{Y} + \frac{N-n}{Nn}\bar{Y}\left[\frac{s_x^2}{\bar{X}^2} - \frac{\mathrm{cov}\left(x;y\right)}{\bar{X}\bar{Y}}\right]$$

Biais en 1/n

Biais nul si la droite de régression passe par 0.

- Erreur quadratique moyenne

$$E(\bar{y}_q - \bar{Y})^2 = \frac{N-n}{Nn}\left(s_y^2 - 2\frac{\bar{Y}}{\bar{X}} s_{xy} + \left(\frac{\bar{Y}}{\bar{X}}\right)^2 s_x^2 \right) \quad \text{estimé par} \quad \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n} z_i^2$$

Avec $z_i = y_i - r x_i$ où $r = \frac{\bar{y}}{\bar{x}}$

Complément: estimation d'un ratio

- Exemple: tirage de n exploitations agricoles (élevage): X_i nombre de vaches, Y_i production
- Rendement par vache: $R = \frac{\overline{Y}}{\overline{X}}$ estimé par $r = \frac{\overline{y}}{\overline{x}}$
- Rapport de deux variables aléatoires
- Développement limité

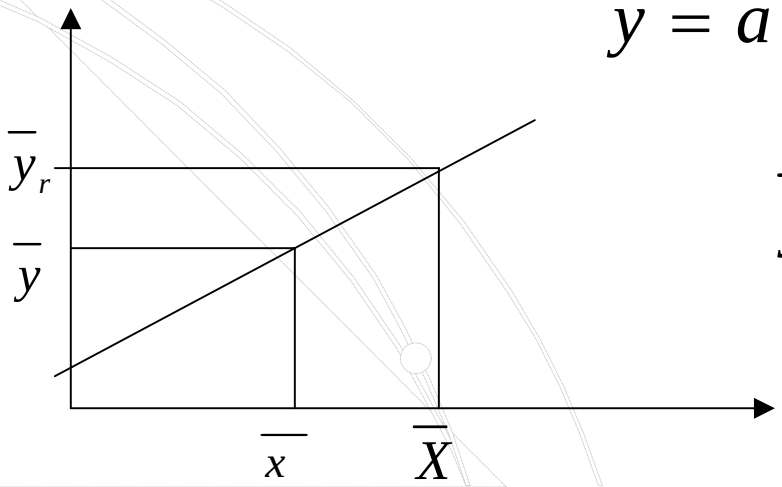
$$E(r) \simeq R + \frac{N-n}{Nn} R \left(\frac{s_x^2}{\overline{X}^2} - \frac{s_{xy}}{\overline{X}\overline{Y}} \right)$$

Estimation par la régression

On connaît pour chaque individu de l'échantillon une variable de contrôle x_i et aussi la valeur moyenne $\overline{X}$ sur la population .

Hypothèse:

$$y = a + bx$$



$$\overline{y}_r = \overline{y} + b\left(\overline{X} - \overline{x}\right)$$

Post-stratification pour une variable numérique

$$\hat{T}_{post} = \sum N_h \bar{y}_h \qquad \hat{\bar{Y}}_{post} = \frac{1}{N} \sum N_h \bar{y}_h$$

Exemple: enquête concernant les revenus

X=classe d'âge; Y=revenu

<20	21-35	36-50	>50
15%	30%	30%	25%
6000	9000	15 000	12 000

$\bar{y} = 11100$

On sait que les vraies proportions sont:

20 35 30 15 $\hat{\bar{y}}_{post} = 10650$

Post-stratification pour une variable numérique

$$V\left(\hat{\bar{Y}}_{post}\right) = \underbrace{V\left[E\left(\hat{\bar{Y}}_{post}/n_h\right)\right]}_{0} + E\left[V\left(\hat{\bar{Y}}_{post}/n_h\right)\right]$$

Conditionnellement aux n_h:

$$\sum\left(\frac{N_h}{N}\right)^2 V\left(\bar{y}_h\right) = \sum\left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h n_h} S_h^2 = \sum\left(\frac{N_h}{N}\right)^2 S_h^2 \frac{1}{n_h} - \frac{1}{N}\sum\left(\frac{N_h}{N}\right) S_h^2$$

En prenant l'espérance:

$$\sum\left(\frac{N_h}{N}\right)^2 S_h^2 E\left(\frac{1}{n_h}\right) - \frac{1}{N}\sum\left(\frac{N_h}{N}\right) S_h^2$$

Calcul de $E\left(\frac{1}{n_h}\right)$

$$P_h = \frac{N_h}{N} \qquad p_h = \frac{n_h}{n}$$

$$n_h = n\frac{n_h}{n} = np_h = n\left(p_h - P_h + P_h\right) = nP_h\left(1 + \frac{p_h - P_h}{P_h}\right)$$

Développement limité

$$\frac{1}{n_h} = \frac{1}{n\,P_h} \times \frac{1}{1 + \underbrace{\frac{p_h - P_h}{P_h}}_{\varepsilon}}$$

$$\frac{1}{n_h} \simeq \frac{1}{nP_h}\left[1 - \varepsilon + \varepsilon^2\right] = \frac{1}{nP_h}\left[1 - \frac{p_h - P_h}{P_h} + \left(\frac{p_h - P_h}{P_h}\right)^2\right]$$

En prenant l'espérance :

$$E(p_h) = P_h \qquad V(p_h) = \frac{N-n}{Nn} P_h (1-P_h)$$

$$E\left(\frac{1}{n_h}\right) = \frac{1}{nP_h}\left[1 + \frac{N-n}{Nn^2} \times \frac{Q_h}{P_h}\right]$$

$$V(\bar{y}_{post}) = \sum P_h^2 S_h^2 \left[\frac{1}{nP_h} + \frac{N-n}{Nn^2} \times \frac{Q_h}{P_h}\right] - \frac{1}{N}\sum P_h S_h^2$$

$$= \frac{N-n}{Nn}\sum P_h S_h^2 + \frac{1}{n}\frac{N-n}{Nn}\sum Q_h S_h^2$$

$$V(\hat{\bar{Y}}_{post}) \simeq \left(\frac{1-f}{n}\right)\sum \frac{N_h}{N} S_h^2 + \frac{1-f}{n^2}\sum\left(1 - \frac{N_h}{N}\right) S_h^2$$

Poids de redressement

$$\hat{\bar{Y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \frac{1}{n_h} \sum_{i_h}^{n_h} y_{i_h} = \sum_{h=1}^{H} \sum_{i_h}^{n_h} \frac{N_h}{N n_h} y_{i_h}$$

Poids de redressement: $\frac{N_h}{N n_h}$

- 1= somme des poids de redressement sur les *n* unités de l'échantillon
- Estimation de $\bar{Y}$ comme moyenne pondérée des valeurs observées.
- Ne pas confondre poids de redressement et poids d'échantillonnage (probabilités d'inclusion).

Post-stratification; redressement sur critère qualitatif

Exemple:

on veut estimer le taux de fréquentation du cinéma.

La fréquentation du cinéma est liée à la possession de TV.

On sait que $\tau_{télé} = 80\%$.

Post-stratification; redressement sur critère qualitatif

Télé \ Cinéma	Oui	Non	Total
Oui	20	680	700
Non	80	220	300
Total	100	900	

(800) X 8/7

(200) x 2/3

Après redressement:

Télé \ Cinéma	Oui	Non	Total
Oui	23	777	800
Non	53	147	200
Total	76	924	

Généralisation: calage sur marges

- Redressement sur plusieurs critères
  - Méthode itérative de Deming et Stephan (RAS)

    On ajuste alternativement sur chaque marge (succession de règles de 3)
  - Macro CALMAR de l'INSEE

1000 individus ont été interrogés. La répartition par sexe et profession est la suivante

	P1	*P2*	*P3*	*Total*
H	300	100	200	*600*
F	100	150	150	*400*
Total	*400*	*250*	*150*	*1000*

Vraies marges 500 et 500 pour le sexe et 350,300, 350 pour la profession.

Une première règle de 3 permet d'obtenir les marges souhaitées pour le sexe : on multiplie la première ligne par 500/600 et la deuxième ligne par 500/400

	P1	*P2*	*P3*	*Total*
H	250	83	167	*500*
F	125	187.5	187.5	*500*
Total	*375*	*270.5*	*354.5*	*1000*

On redresse ensuite en colonne pour ajuster les effectifs marginaux de la variable profession, ce qui change les marges en ligne :

	P1	*P2*	*P3*	*Total*
H	233	92	165	*490*
F	117	208	185	*510*
Total	*350*	*300*	*350*	*1000*

Puis en ligne :

	P1	*P2*	*P3*	*Total*
H	238	94	168	*500*
F	115	204	181	*500*
Total	*353*	*298*	*349*	*1000*

En l'absence de cases vides, l'algorithme converge rapidement et donne les poids de redressement à appliquer à chaque case. Ainsi à la quatrième itération (très proche du résultat souhaité) , les 300 individus H et P1 ont chacun un poids de 0.236. La somme des poids de redressement des 1000 individus vaut 1000.

	P1	*P2*	*P3*	Total
H	236	95	168	499
F	114	205	182	501
Total	*350*	*300*	*350*	1000

- Pour avoir une bonne post-stratification
  - Variable de redressement bien corrélée avec Y
  - n grand
  - $(N-N_h)/N$ petit donc grandes strates
  - Effectifs N_h ou poids des post-strates connus
- Mais:
  - ne pas utiliser que des variables socio-démographiques
  - redressements sur trop de critères dangereux