Chapitre 7

Estimation par régression

Dans ce chapitre nous étudions l'estimation par régression.

7.1 Introduction

On dispose souvent, quand on doit faire un sondage, d'une *information auxiliaire* sous la forme d'une ou plusieurs variables x, connues pour tous les individus de la population U et corrélées avec la variable d'étude y. Les techniques d'estimation par régression servent à incorporer cette information auxiliaire dans les estimateurs par sondage. L'information auxiliaire peut être, par exemple : la variable d'étude connue à une époque antérieure, des variables proches de la variable d'étude mais peu coûteuses à observer. On verra que souvent, on n'utilise que la somme sur la population d'une variable auxiliaire et ses valeurs sur l'échantillon et non sur toute la population.

On a utilisé une telle information dans le plan stratifié et dans l'estimation post-stratifiée. En effet, définissons $x_{hk} = 1$ si $k \in U_h, = 0$ sinon, alors les effectifs des strates sont $N_h = \sum_U x_{hk}$. Dans le plan stratifié on utilise cette information pour définir le plan de sondage alors que dans la post-stratification on l'utilise pour corriger l'estimateur de H-T obtenu sans cette information. C'est cette dernière idée qui est mise en œuvre dans l'estimation par régression.

Dans le présent chapitre, on étudie d'abord l'estimation du total par ratio puis par différence. L'estimateur par différence dépend de paramètres rarement connus. On définit parallèlement un *modèle de superpopulation* ; c'est un modèle de régression qui suppose que la population finie U qui nous intéresse est elle-même tirée d'une population infinie. Ce cadre permet d'étendre l'estimateur par différence quand ses paramètres sont inconnus. De plus on verra que suivant le modèle de superpopulation adopté, on peut obtenir l'estimateur par ratio ou l'estimateur poststratifié.

7.2 Estimation par ratio

7.2.1 Définition

On s'intéresse à l'estimation du total t_y d'une variable d'étude y. On suppose que l'on dispose d'une variable auxiliaire x pour laquelle on connait le total t_x sur toute la population et qui est bien corrélée avec la variable d'étude. On définit l'estimateur par ratio du total $\widehat{t}_{y\mathtt{ra}}$ par :

$$\widehat{t}_{y\mathtt{ra}} = \hat{t}_{y\pi} \times \frac{t_x}{\hat{t}_{x\pi}}$$

On peut interpréter cette définition comme "une règle de trois" qui permet d'ajuster l'estimateur par les valeurs dilatés $\hat{t}_{y\pi}$ d'un coefficient multiplicatif qui tient compte de la qualité de l'estimation du total

par l'estimateur de H-T pour la variable x pour l'échantillon tiré. On peut aussi écrire :

$$\widehat{t}_{y\mathtt{ra}} = \widehat{R} \times t_x.$$

Cette dernière égalité définit l'estimateur par ratio comme l'estimateur d'un ratio multiplié par une constante t_x (non aléatoire). A partir des formules de variances de $\widehat{R}$ du chapitre 6, on déduit facilement la variance de l'estimateur par ratio.

7.2.2 Propriétés de l'estimateur par ratio

1. $\widehat{t}_{y\mathtt{ra}}$ est approximativement sans biais pour t_y
2. d'après les résultats (5) et (6) du chapitre 6, dans le cas d'un plan SI, une approximation de sa variance est
$$\mathsf{var}(\widehat{t}_{y\mathtt{ra}}) = t_{xU}^2 \mathsf{var}(\widehat{R}) \simeq N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-Rx,U}. \tag{7.1}$$
mais
$$\widehat{\mathsf{var}}(\hat{R}) = \frac{1}{\overline{x}_s^2} \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}x,s}.$$
3. On estime $\mathsf{var}(\widehat{t}_{y\mathtt{ra}})$ par
$$\widehat{\mathsf{var}}(\widehat{t}_{y\mathtt{ra}}) = N^2 \frac{\overline{x}_U^2}{\overline{x}_s^2} \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}x,s}. \tag{7.2}$$

Bien noter le facteur $\dfrac{\overline{x}_U^2}{\overline{x}_s^2}$ qui arrive quand on passe de l'estimateur à l'estimation. On rencontre parfois l'estimateur de variance :

$$N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{y-\hat{R}x,s} \tag{7.3}$$

obtenu par une subsitution directe dans (7.1). Si $\overline{x}_U \simeq \overline{x}_s$, les deux estimateurs sont proches. Observons que (7.3) est l'estimation de la variance du total des résidus $y_k - \widehat{R}x_k$.

7.2.3 Efficacité de l'estimateur par ratio pour le plan SI

Examinons le rapport : variance de l'estimateur par ratio sur variance de l'estimateur par les valeurs dilatées dans le plan SI :

$$\frac{S^2_{yU} - 2\widehat{R}\, S_{yx,U} + R^2\, S^2_{xU}}{S^2_{yU}} = 1 - 2R\, \frac{S_{yx,U}}{S^2_{yU}} + R^2\, \frac{S^2_{xU}}{S^2_{yU}}.$$

Dans cette expression $R = t_y/t_x = \overline{y}_U/\overline{x}_U$. Introduisons les coefficients de variation : $\mathrm{cv}(yU) = S_{t=yU}/\overline{y}_U$ et $\mathrm{cv}(xU)$, alors on voit que ce rapport est < 1 si le coefficient de corrélation entre y et x, ρ vérifie

$$\rho \geq \frac{1}{2} \frac{\mathrm{cv}(xU)}{\mathrm{cv}(yU)}$$

L'amélioration de l'estimation du total quand on passe à l'estimation par ratio, dépend de la situation du coefficient de corrélation par rapport au rapport des variabilités relatives de x et y.

7.3 Estimation par différence

7.3.1 Définition et propriétés

On veut estimer le total d'une variable y : $t_{yU} = \sum_U y_k$. Supposons que la variable d'étude y prenne des valeurs très proches d'une variable y^0 connue pour chaque individu de la population U, qu'on appellera un *proxy* :

$$y_k \simeq y_k^0 \; \forall k \in U$$

Posons $D_k = y_k - y_k^0$ et écrivons :

$$t_{yU} = \sum\nolimits_U y_k = \sum\nolimits_U y_k^0 + \sum\nolimits_U D_k.$$

La somme, $\sum_U y_k^0$, n'est pas aléatoire et est connue. Il reste donc à estimer $\sum_U D_k$. Les D_k fluctuent beaucoup moins que les y_k. On peut maintenant donner la définition d'un estimateur par différence. Etant donné un plan de sondage, de probabilités d'inclusion : $\pi_k,\ \pi_{kl}$, on appelle, estimateur par différence de t_{yU}, la quantité :

$$\widehat{t}_{y,\text{diff}} = \sum\nolimits_U y_k^0 + \sum\nolimits_s \check{D}_k$$

où, $\check{D}_k = \frac{y_k - y_k^0}{\pi_k}$.
On voit immédiatement que c'est un estimateur sans biais. Sa variance est celle de sa composante aléatoire :

$$\mathsf{var}(\widehat{t}_{y,\text{diff}}) = \mathsf{var}\left(\sum\nolimits_s \check{D}_k\right).$$

Elle est d'autant plus faible que y_k^0 est proche de y_k. En somme la variance de l'estimateur par différence du total est la variance de l'estimateur du total des "résidus" D_k, situation déjà observée sur (7.3).

Cas particuliers

1. Plan de taille fixe. On a

$$\mathsf{var}(\widehat{t}_{y,\text{diff}}) = -\frac{1}{2}\sum\sum\nolimits_U \Delta_{kl}(\check{D}_k - \check{D}_l)^2$$

2. Plan SI, taux de sondage : $f = n/N$. Notant $x_k = y_k^0$ on obtient :

$$\mathsf{var}_{\text{SI}}(\widehat{t}_{y,\text{diff}}) = N^2\frac{1-f}{n}\{S_{yU}^2 + S_{xU}^2 - 2S_{xyU}\}$$

Efficacité de l'estimateur par différence dans le plan SI

Introduisons le coefficient de corrélation sur population finie :

$$r = \frac{S_{xyU}}{S_{xU}S_{yU}}$$

Il est facile de vérifier que le gain en terme de variance de l'estimateur par différence est :

$$\mathsf{var}_{\text{SI}}(\widehat{t}_{y,\text{diff}})/\mathsf{var}_{\text{SI}}(\widehat{t}_{y,\pi}) = 1 + (\frac{S_{xU}}{S_{yU}})^2 - 2r\frac{S_{xU}}{S_{yU}}$$

et que cet estimateur apporte un gain sur l'estimateur par les valeurs dilatées, seulement si

$$r > \frac{1}{2}\frac{S_{xU}}{S_{yU}}$$

7.3.2 Autre point de vue sur l'estimation par différence

En vue de la suite du chapitre, supposons que y_k^0 est une combinaison de variables auxiliaires, à coefficients connus :

$$y_k^0 = A'\mathbf{x}_k$$

où $\mathbf{x}_k$ est donc la matrice colonne des J variables auxiliaires pour k et A une matrice $J \times 1$. Avec des notations désormais classiques, on voit que :

$$\widehat{t}_{y\text{diff}} = \widehat{t}_{y,\pi} + \sum_{j=1}^{J} A_j(t_{x_j} - \widehat{t}_{x_j,\pi})$$

Ainsi, on peut considérer l'estimateur par différence comme l'estimateur usuel par les valeurs dilatées corrigé d'après la différence entre la valeur exacte du total de la variable auxiliaire et son estimation par les valeurs dilatées. Supposons en particulier qu'il n'y a qu'une variable auxiliaire :

$$\widehat{t}_{y\text{diff}} = \widehat{t}_{y,\pi} + A_1(t_{x_1} - \widehat{t}_{x_1,\pi})$$

avec $A_1 > 0$. On voit que si un échantillon surestime t_x et donc t_y, la correction $A_1(t_{x_1} - \widehat{t}_{x_1,\pi})$ est négative et vient compenser la surestimation de l'échantillonnage.
Le plus souvent, la matrice A n'est pas connue. On doit donc la remplacer par une estimation, on obtient alors un *estimateur par régression*. De nombreuses façons de choisir A existent, chacune répond à un critère particulier d'optimalité. Dans ce cours on estime A par moindres carrés. On va précisément aborder ce point de vue en considérant que la population finie U est elle-même un échantillon au sens de la statistique inférentielle habituelle, prélevé dans une population infinie.

7.4 Estimateur par régression

7.4.1 Définition

Supposons que la population finie U est elle-même obtenue par des tirages indépendants dans une population infinie et définissons le *modèle de superpopulation* :

$$\begin{aligned} E_\xi(y_k) &= \sum_{j=1}^{J} \beta_j x_{jk} = \mathbf{x}_k'\beta, \\ V_\xi(y_k) &= \sigma_k^2 \end{aligned} \qquad k \in U \qquad (7.4)$$

L'indice ξ fait référence à la loi sur la population parente infinie.

Le problème qui nous intéresse est toujours d'estimer le total de y sur U. Quel que soit le plan de sondage mis en œuvre, on observe y et les x pour chaque élément de l'échantillon et on connaît les valeurs des x pour tout U.

On s'inspire de l'estimateur par différence. Il faut donc estimer le coefficient A apparaissant dans cette méthode. La méthode des moindres carrés pondérés suggère un tel estimateur. Ceci observé, la construction se déroule naturellement.

Indication pratique. Si on a oublié les notations matricielles dans la méthode des moindres carrés, une façon simple de vérifier les calculs déroulés ci-dessous consiste à les détailler entièrement sur une population de 3 individus et avec 2 variables explicatives.

Si l'échantillon au sens de la superpopulation, c'est-à-dire U tout entier, était disponible, on estimerait β par :

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{t} \tag{7.5}$$

$$\text{où } \mathbf{T} = \sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2}, \text{ et } \mathbf{t} = \sum_U \frac{\mathbf{x}_k y_k}{\sigma_k^2} \tag{7.6}$$

dans ces expressions, $\mathbf{T}$ est l'habituel : $(\mathbf{X}'W\mathbf{X})^{-1}$ de la méthode des moindres carrés pondérés, $\mathbf{t}$ est $\mathbf{X}'Wy$ où W est la matrice diagonale des poids $1/\sigma_k^2$.

Mais $\mathbf{T}$ et $\mathbf{t}$ sont des matrices de totaux sur U, donc non calculables. On les estime à partir de s.

$$\widehat{T} = \sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k} \text{ et } \widehat{\mathbf{t}} = \sum_s \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k} \tag{7.7}$$

puis :

$$\widehat{\mathbf{B}} = \widehat{T}^{-1}\widehat{\mathbf{t}} \tag{7.8}$$

Commentons. $\mathbf{T}$ et t sont des matrices de totaux sur U. Chacun est estimé, comme dans l'estimation d'un ratio, par son estimateur par les valeurs dilatées. Ce qui donne l'estimateur, $\widehat{\mathbf{B}}$, au sens du plan de sondage (= design based), de l'estimateur, $\mathbf{B}$, au sens du modèle ξ (=model based), de β, qu'on peut aussi écrire :

$$\widehat{\mathbf{B}} = \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k} \right)^{-1} \sum_s \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k} \tag{7.9}$$

Ou encore, l'estimateur des coefficients de régression s'obtient à partir de l'estimateur classique des moindres carrés pondérés de β dans lequel on remplace chaque composante par son estimateur par les valeurs dilatées.

L'estimateur du total par régression est alors défini en les mêmes termes que l'estimateur par différence.

$$\widehat{t}_{yr} := \widehat{t}_{y\pi} + (t_{\mathbf{x}U} - \widehat{t}_{\mathbf{x}\pi})'\widehat{\mathbf{B}} = \widehat{t}_{y\pi} + \sum_{j=1}^{J} (t_{xj} - \widehat{t}_{x_j,\pi})\widehat{B}_j. \tag{7.10}$$

En remplaçant dans cette expression, $\widehat{\mathbf{B}}$ par son écriture (7.9), on vérifie facilement que cet estimateur s'écrit :

$$\widehat{t}_{yr} = \sum_s g_{ks} \check{y}_k$$

où

$$g_{ks} = 1 + (t_{\mathbf{x}U} - \widehat{t}_{\mathbf{x}\pi})'\widehat{T}^{-1} \frac{\mathbf{x}_k}{\sigma_k^2} \tag{7.11}$$

L'estimateur par régression est donc un estimateur pondéré mais les poids dépendent de l'échantillon, au contraire de ce qui se passe avec l'estimateur par les valeurs dilatées.

Illustration. Nous détaillons maintenant les formules quand il y a 2 variables auxiliaires : x_1 et x_2.

$$\mathbf{x}_k = \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix}, \ \mathbf{T} = \sum_U \frac{1}{\sigma_k^2} \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} \begin{bmatrix} x_{1k} & x_{2k} \end{bmatrix} = \begin{bmatrix} \sum_U \frac{1}{\sigma_k^2} x_{1k}^2 & \sum_U \frac{1}{\sigma_k^2} x_{1k}x_{2k} \\ \sum_U \frac{1}{\sigma_k^2} x_{2k}x_{1k} & \sum_U \frac{1}{\sigma_k^2} x_{2k}^2 \end{bmatrix}.$$

$$\mathbf{t} = \sum_U \frac{1}{\sigma_k^2} \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} y_k = \begin{bmatrix} \sum_U \frac{1}{\sigma_k^2} x_{1k} y_k \\ \sum_U \frac{1}{\sigma_k^2} x_{2k} y_k \end{bmatrix}.$$

On construit ensuite les estimateurs par les valeurs dilatées : $\widehat{\mathbf{T}}, \widehat{\mathbf{t}}$.

$$\widehat{\mathbf{T}} = \sum_s \frac{1}{\pi_k \sigma_k^2} \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} \begin{bmatrix} x_{1k} & x_{2k} \end{bmatrix} = \begin{bmatrix} \sum_s \frac{1}{\pi_k \sigma_k^2} x_{1k}^2 & \sum_s \frac{1}{\pi_k \sigma_k^2} x_{1k} x_{2k} \\ \sum_s \frac{1}{\pi_k \sigma_k^2} x_{2k} x_{1k} & \sum_s \frac{1}{\pi_k \sigma_k^2} x_{2k}^2 \end{bmatrix}.$$

$$\widehat{\mathbf{t}} = \sum_s \frac{1}{\pi_k \sigma_k^2} \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} y_k = \begin{bmatrix} \sum_s \frac{1}{\pi_k \sigma_k^2} x_{1k} y_k \\ \sum_s \frac{1}{\pi_k \sigma_k^2} x_{2k} y_k \end{bmatrix}.$$

Remarque. Considérons le cas où il y a une constante dans la régression, $x_{1k} = 1 \forall k$ et où la variance est constante : $\sigma_k^2 = \sigma^2$. Dans ce cas $\sum_U \frac{1}{\sigma_k^2} x_{1k}^2 = 1/\sigma^2 N$ est connu et on peut être tenté de remplacer dans $\widehat{\mathbf{T}}$, son estimation par cette valeur connue. Ceci est incorrect. Il faut traiter tous les éléments de $\mathbf{T}$ et $\mathbf{t}$ de la même façon.

Cas particulier.
Supposons $J = 1$ et $v_k = x_k$ alors, $\widehat{B} = \widehat{R}$ et on est ramené à l'estimateur par Ratio.

7.4.2 Exercice 2

Dans une certaine commune, il y a $N = 1000$ propriétés immobilières. Elles occupent une surface d'environ 31.5 ha de terrain sur la commune. Cinq d'entre elles ont été vendues le mois dernier.

Les 5 transactions

Propriété	Surface de la parcelle en m^2	Prix de vente en kF
1	130	935
2	255	1170
3	510	1920
4	340	1500
5	450	1900

Estimons la valeur totale de la propriété immobilière dans la commune.

Réponse. Commençons par modéliser ce problème. Supposons d'abord que les propriétés vendues le mois précédent sont assimilables à un échantillon tiré suivant un plan SI. Appelons y_k la valeur de la propriété k, on veut estimer $t_y = \sum_U y_k$, où U est la population des propriétés immobilières. Comme l'échantillon est très petit, le diagramme de dispersion des 5 points collectés ne peut pas suggérer un modèle de variance plutôt qu'un autre. D'autre part, vu la nature du problème, l'ordonnée à l'origine est non nulle et l'on peut penser, que comme dans la plupart des problèmes où une variable de taille intervient, la variance de y est une fonction croissante de x. Supposons

$$E_\xi(y_k) = \beta_0 + \beta_1 x_k$$
$$V_\xi(y_k) = \sigma^2 x_k$$

$\widehat{\mathbf{B}} = \widehat{T}^{-1}\widehat{\mathbf{t}}$ s'obtient ainsi :

$$\mathbf{X} = \begin{bmatrix} 1 & 130 \\ 1 & 255 \\ 1 & 510 \\ 1 & 340 \\ 1 & 450 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 935 \\ 1170 \\ 1920 \\ 1500 \\ 1900 \end{bmatrix}.$$

La matrice diagonale des poids est au facteur près $1/(\sigma^2\pi_k)$ constant dans le plan SI :

$$W = \begin{bmatrix} 0.0076923 & 0 & 0 & 0 & 0 \\ 0 & 0.0039216 & 0 & 0 & 0 \\ 0 & 0 & 0.0019608 & 0 & 0 \\ 0 & 0 & 0 & 0.0029412 & 0 \\ 0 & 0 & 0 & 0 & 0.0022222 \end{bmatrix}.$$

D'où :

$$(1000/5)\ \sigma^2\ \widehat{\mathbf{T}} = \begin{bmatrix} 3.7476 & 1000 \\ 1000 & 337000 \end{bmatrix}, \qquad (1000/5)\ \sigma^2\ \widehat{\mathbf{t}} = \begin{bmatrix} 550.2306 \\ 2.7738 \end{bmatrix}$$

et

$$\widehat{\mathbf{B}} = \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}} = \begin{bmatrix} 550.2306 \\ 2.7738 \end{bmatrix}.$$

D'autre part $t_{xU} = 315000$, $\widehat{t}_{y\pi} = N\overline{y}_s = 1485000$, enfin

$$\widehat{t}_{yr} = \widehat{t}_{y\pi} + (N-N)\widehat{b}_1 + (t_{xU} - \widehat{t}_{x\pi})\widehat{b}_2 = 1423976$$

```
se calcule ainsi (en R)  : a <- matrix( c(130 , 935,  255 , 1170,
510 , 1920,  340  , 1500,
     450 ,  1900), nrow= 5, byrow=T, ncol=2)
x <- a[,1] y <- a[,2] colnames(a) <- c("Surf","Prix")

X <- matrix(c(rep(1,5),x) ,nrow= 5,  ncol=2, byrow=F) X
     [,1] [,2]
[1,]    1  130 [2,]    1  255 [3,]    1  510 [4,]    1  340 [5,]
1  450

w <- diag(1/x) w
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.0076923 0.0000000 0.0000000 0.0000000 0.0000000 [2,]
0.0000000 0.0039216 0.0000000 0.0000000 0.0000000 [3,] 0.0000000
0.0000000 0.0019608 0.0000000 0.0000000 [4,] 0.0000000 0.0000000
0.0000000 0.0029412 0.0000000 [5,] 0.0000000 0.0000000 0.0000000
0.0000000 0.0022222

n <- 5
 N <- 1000
 T <- (N/n)*t(X) %*% w %*% X
 T
          [,1]   [,2]
[1,]    3.7476   1000 [2,] 1000.0000 337000
 t <- (N/n)*t(X) %*% w %*% y
 B <- solve(T,t)
 B
          [,1]
[1,] 550.2306 [2,]   2.7738
```

et l'estimation (7.10) s'obtient par :

```
>  txu <- 315000
> ty.ht <- N *mean(y)
> ty.ht
[1] 1485000
> ty.reg <- ty.ht + (N - N)*B[1,1] + (txu - N * mean(x))* B[2,1]
> ty.reg
[1] 1423976
```

Remarques.

- Observons que (7.9) est l'estimateur obtenu par régression pondérée classique de y sur les x pour l'échantillon s avec les poids $w_k = 1/(\sigma_k^2 \pi_k)$. Il est donc facile de calculer ce coefficient de régression avec n'importe quel logiciel d'analyse statistique, pour un plan simple.
- Une procédure `surveyreg` dans SAS, fait ce travail pour différents plans et calcule également les estimations de variance que nous voyons maintenant.

7.4.3 Variance de l'estimateur par régression

On obtient une expression approchée de l'estimateur par régression $\widehat{t}_{yr}$ en développant à l'ordre 1 l'expression de $\widehat{t}_{yr}$. On peut voir Särndal et al. pour les détails. L'approximation à l'ordre 1 obtenue est :

$$\widehat{t}_{y,r_0} = \widehat{t}_{y,\pi} + (t_{\mathbf{x}} - \widehat{t}_{\mathbf{x},\pi})'\mathbf{B}$$

On introduit des y ajustés théoriques (car non calculables d'après le seul échantillon s) ainsi que des résidus théoriques :

$$y_k^0 = \mathbf{x}_k'\mathbf{B},\ k \in U \tag{7.12}$$

$$E_k = y_k - y_k^0,\ k \in U. \tag{7.13}$$

On définit également des y ajustés empiriques et des résidus empiriques :

$$\widehat{y}_k = \mathbf{x}_k'\widehat{\mathbf{B}},\ \ k \in U$$

$$e_{k,s} = y_k - \widehat{y}_k,\ \ k \in s$$

Avec ces notations, on montre que la variance de l'estimateur linéarisé $\widehat{t}_{y,r_0}$ qui est la variance approchée de $\widehat{t}_{y,r}$ vaut :

$$\mathsf{var}_{\text{app}}(\widehat{t}_{y,r}) = \mathsf{var}(\widehat{t}_{y,r_0}) = \sum\sum\nolimits_U \Delta_{kl}\check{E}_k\check{E}_l$$

$$\text{approchée par : } \widehat{\mathsf{var}}(\widehat{t}_{y,r}) = \sum\sum\nolimits_s \check{\Delta}_{kl}(g_{ks}\check{e}_{ks})(g_{ls}\check{e}_{ls}) \tag{7.14}$$

En particulier, pour un plan SI(N,n) :

$$\widehat{\mathsf{var}}(\widehat{t}_{y,r}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right) S^2_{g_s e_s\, s} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum\nolimits_s (g_{ks}e_{ks} - \overline{g_s e_s})^2 \tag{7.15}$$

Commentaires pratiques. Observons d'abord que cette expression fait intervenir

1. des covariances d'indicatrices d'inclusion, des probabilités d'inclusion d'ordre 2, ceci comme pour la variance de n'importe quel estimateur de Horvitz-Thompson,

2. des résidus de régressions linéaires pondérées, parfois difficiles à calculer,
3. des coefficients g_{ks}.

Tout ceci est lourd à mettre en œuvre.

1. Une première simplification consiste à considérer que $g_{ks} \simeq 1$. Alors (7.14) devient

$$\widehat{\mathsf{var}}(\widehat{t}_{y,r}) = \sum \sum_s \check{\Delta}_{kl} \check{e}_{ks} \check{e}_{ls},$$

   qui n'est autre que l'estimateur de la variance de l'estimateur du total des résidus de la régression. Par exemple, pour le plan SI, cette simplification donnera comme estimation :

$$\widehat{\mathsf{var}}(\widehat{t}_{y,r}) = N^2(\frac{1}{n} - \frac{1}{N})S^2_{e_s\ s} = N^2(\frac{1}{n} - \frac{1}{N})\frac{1}{n-1}\sum_s (e_{ks} - \overline{e_s})^2 \tag{7.16}$$

2. D'autre part on voit qu'à g_{ks} près, l'estimateur par régression est un estimateur par différence utilisant un proxy particulier pour y. Les procédures de régression de beaucoup de logiciels ne permettent pas de calculer exactement ce proxy, de plus d'autres combinaisons linéaires des x peuvent être légitimement choisies comme proxy. L'important est de disposer des résidus d'une régression pondérée de y sur les x, bien choisie par rapport aux données.

Enfin certains modèles de superpopulation débouchent sur des calculs plus simples. C'est ce que nous verrons au paragraphe suivant.

7.4.4 Exercice 2 (suite)

Les résidus e_{ks} sont :

```
24.17587  -87.54869  -44.86679    6.67861  101.56100
```

$S^2_{e_s\ s} = 5155.379$.

Les g_{ks} sont :

```
0.7654873 0.8484436 0.8915808 0.8700122 0.8858292
```

On note qu'ils ne sont pas trop éloignés de 1.

Les $g_{ks}e_{ks}$ sont :

```
18.506 -74.28 -40.002 5.8105 89.966
```

et $S^2_{g_s e_s\ s} = 3896.95$. D'où l'estimation de la variance du total :

$$\widehat{\mathsf{var}}(\widehat{t}_{yr}) = 1000^2 \times (1/5 - 1/1000) \times 3896.95 = 775493013$$

L'estimation simplifiée (7.16) est :

$$\widehat{\mathsf{var}}_{\mathtt{simpl}}(\widehat{t}_{yr}) = 1000^2 \times (1/5 - 1/1000) \times 5155.379 = 1025920368.$$

D'autre part on trouve $S^2_{ys} = 190850$ et la variance estimée de l'estimateur de H.-T. est :

$\widehat{\mathsf{var}}(\widehat{t}_{y\pi}) = 1000^2 \times (1/5 - 1/1000) \times 190850 = 37979150000.$

Pour récapituler écrivons les coefficients de variations des estimateurs :

$$\mathrm{cv}(\widehat{t}_{y\pi}) = \frac{\widehat{\mathsf{var}}(\widehat{t}_{y\pi})^{.5}}{\widehat{t}_{y\pi}} = \frac{37979150000^{.5}}{1485000} = 0.1312339$$

$$\mathrm{cv}(\widehat{t}_{yr}) = \frac{\widehat{\mathsf{var}}(\widehat{t}_{yr})^{.5}}{\widehat{t}_{yr}} = \frac{775493013^{(}.5)}{1423976} = 0.01955627$$

$$\mathrm{cv}_{\mathtt{simpl}}(\widehat{t}_{yr}) = \frac{\widehat{\mathsf{var}}_{\mathtt{simpl}}(\widehat{t}_{yr})^{.5}}{\widehat{t}_{yr}} = \frac{1025920368^{.5}}{1423976} = 0.02249334.$$

On observe la plus grande précision de l'estimateur par régression.

7.4.5 Compléments

Trois écritures équivalentes de l'estimateur par régression.

$$\begin{aligned} \widehat{t}_{yr} &= \sum\nolimits_s g_{ks}\check{y}_k && \text{(e1)} \\ &= \sum\nolimits_U \widehat{y}_k + \sum\nolimits_s \check{e}_{ks} && \text{(e2)} \\ &= \sum\nolimits_U y_k^0 + \sum\nolimits_s g_{ks}\check{E}_k && \text{(e3)} \end{aligned}$$

Exercice. Montrer l'équivalence des 3 expressions de l'estimateur par régression. Indication : Montrer que l'expression générale (7.10) implique (e1), (e2) et (e3).
Dans certains cas, la somme des résidus est nulle, et (e2) se réduit à $\widehat{t}_{yr} = \sum_U \widehat{y}_k$, ce qui simplifie sensiblement les calculs.

Résultat 1 (SSW 6-5-1) *Etant donné un plan de sondage, une condition suffisante pour que*

$$\sum\nolimits_s \check{e}_{ks} = 0 \forall s \text{ parmi les échantillons possibles de ce plan}$$

est qu'il existe $\lambda, \quad J \times 1$*, indépendant de* k *tel que :*

$$\sigma_k^2 = \lambda' \mathbf{x}_k \forall k \in U$$

Exemples d'une telle situation.
- σ_k^2 est constante et le modèle ne passe pas par l'origine.
- σ_k^2 proportionnelle à un des x

$$\sigma_k^2 \propto x_{jk} \quad \exists j \in \{1, \cdots, J\} \quad \forall k \in U$$

- σ_k^2 proportionnelle à une combinaison linéaire des x :

$$\sigma_k^2 \propto a' \mathbf{x}_k \quad \forall k \in U$$

pour un certain $a, \quad J \times 1$ indépendant de k

Comment reconnaître la situation qui convient dans un problème empirique particulier ? Une fois que l'échantillon a été récolté, faire un graphique de y contre chaque x et voir si la dispersion des points autour d'une direction d'ajustement est à peu près constante, ou plus ou moins croissante.

7.5 Cas particuliers

Nous avons vu d'abord l'estimation par ratio puis l'estimation par différence et enfin l'estimation par régression en général. Maintenant nous considérons des modèles élémentaires de régression : moyenne constante et modèle d'ANOVA à un facteur.

7.5.1 Modèle à moyenne constante

Le modèle est :

$$\begin{aligned} E_\xi(y_k) &= \beta, \\ V_\xi(y_k) &= \sigma^2 \end{aligned} \qquad k \in U \quad (7.17)$$

Posons :

$$\widehat{N} = \sum\nolimits_s \frac{1}{\pi_k}.$$

et

$$\tilde{y}_s = \sum\nolimits_s \frac{y_k}{\pi_k} / \widehat{N}.$$

$\tilde{y}_s$ est l'*estimateur de Hajek* de la moyenne. Appliquant le principe général de l'estimation par régression, on obtient :

$$\widehat{\beta} = \tilde{y}_s.$$

L'estimateur du total par régression est donc, nous basant sur l'écriture (e2) :

$$\widehat{t}_{yr} = N\tilde{y}_s + \sum_s \check{e}_{ks}$$

avec

$$e_{ks} = y_k - \tilde{y}_s.$$

Faisant naïvement le calcul ou observant que le modèle à moyenne constante remplit la condition suffisante (1), on obtient : $\sum_s \check{e}_{ks} = 0$. D'où,

$$\widehat{t}_{yr} = N\tilde{y}_s$$

On obtient soit directement, soit par application des formules relatives à l'estimation par ratio, l'approximation de la variance de cet estimateur :

$$\widehat{\mathsf{var}}(\widehat{t}_{yr}) = \left(\frac{N}{\widehat{N}}\right)^2 \sum\sum\nolimits_s \check{\Delta}_{kl}\check{e}_{ks}\check{e}_{ls}.$$

Attention. $\widehat{N}$ est l'estimation de Horvitz-Thompson de $\sum_U x_{1k}$ où $x_{1k} = 1 \forall k$. On rejoint ici la remarque faite quand on a explicité les formules pour deux variables explicatives.

7.5.2 Modèle à moyenne de groupe

Exemple. Supposons que la population soit formée d'établissements industriels et commerciaux et que les groupes soient formés par secteurs d'activité. Pour beaucoup de variables, les établissements d'un même secteur se ressemblent. Cette homogénéité intra groupe peut être exploitée pour obtenir des estimateurs améliorés, par régression. L'information auxiliaire est la taille de chaque groupe ou les totaux par groupe de certaines variables auxiliaires.

Les *modèles de groupe* ont un ou plusieurs paramètres associés à chaque groupe particulier. Une fois posés, ils débouchent sur des estimateurs par régression.
Ici, on n'utilise pas les groupes pour stratifier, mais on tire directement dans la population nonpartitionnée et on ne connaît l'appartenance à un groupe que pour les individus sélectionnés après le sondage.

Notations. On note les groupes $U_1, \cdots, U_g, \cdots, U_G$ et N_g est la taille de U_g. On a :

$$\bigcup_{g=1}^{G} U_g = U \text{ et } \sum\nolimits_{g=1}^{G} N_g = N$$

On tire un échantillon s de taille notée n_s. On peut le partitionner en G sous échantillons $s_1, \cdots, s_g, \cdots, s_G$ où $s_g = s \bigcap U_g$, s_g est de taille n_{s_g}. On a également :

$$\bigcup_{g=1}^{G} s_g = s \text{ et } \sum\nolimits_{g=1}^{G} n_{s_g} = n_s$$

Habituellement, n_{s_g} est aléatoire, car le plan porte sur U et U_g fonctionne comme un domaine pour U. Le modèle à moyenne de groupe qui rend compte de l'hypothsèe d'homoscédasticité intra groupe est :

$$\begin{array}{l} E_\xi(y_k) = \beta_g, \\ V_\xi(y_k) = \sigma^2_{0g} \end{array} \qquad k \in U_g, \;\; g = 1, \cdots, G \tag{7.18}$$

C'est un modèle du type : analyse de la variance à un facteur. On se ramène au modèle de régression en introduisant G variables auxiliaires :

$$x_{gk} = \left\{ \begin{array}{lll} 1 & \text{si} & k \in U_g \\ 0 & \text{si} & k \notin U_g \end{array} \right.$$

On obtient ainsi :

$$\mathbf{T} = diag(\frac{N_1}{\sigma^2_{01}}, \cdots, \frac{N_G}{\sigma^2_{0G}}) \; \mathbf{t} = \left[\sum_{U_1} \frac{y_k}{\sigma^2_{01}}, \cdots, \sum_{U_G} \frac{y_k}{\sigma^2_{0G}} \right]'$$

On estime N_g par $\widehat{N}_g = \sum_{s_g} \frac{1}{\pi_k}$ et $\sum_{U_G} y_k$ par $\sum_{s_g} \frac{y_k}{\pi_k}$ (on a rencontré ce genre de calculs dans l'estimation sur un domaine). On peut alors écrire les estimateurs $\widehat{\mathbf{T}}$ et $\widehat{\mathbf{t}}$. On déduit $\widehat{\mathbf{B}}$, matrice $G \times 1$ d'élément g : $\widetilde{y}_{s_g} = \sum_{s_g} \frac{y_k}{\pi_k} / \sum_{s_g} \frac{1}{\pi_k}$ et l'écriture de l'estimateur du total par régression, basée sur (e2) :

$$\widehat{t}_{yr} = \sum_{g=1}^{G} \frac{N_g}{\widehat{N}_g} \sum_{s_g} \check{y}_k + \sum_s \check{e}_{ks}$$

où $e_{ks} = y_k - \widehat{y}_k = y_k - \widetilde{y}_{s_g}$ si $k \in s_g$. Enfin, on observe que le modèle vérifie la condition suffisante pour que $\sum_s \check{e}_{ks} = 0$, et donc

$$\widehat{t}_{yr} = \sum_{g=1}^{G} \frac{N_g}{\widehat{N}_g} \sum_{s_g} \check{y}_k = \sum_{g=1}^{G} N_g \tilde{y}_{s_g}.$$

On reconnaît l'**estimateur post-stratifié**.

Remarque L'information auxiliaire nécessaire se limite à l'effectif de chaque sous-population.

Exercice. Vérifier par le calcul que dans le modèle à moyenne de groupe on a bien : $\sum_s \check{e}_{ks} = 0$.

Cas particulier du plan SI Pour un plan SI, de taux de sondage : $f = n/N$, on a (à vérifier) :

$$\begin{array}{ll} \widehat{N}_g = n_{s_g} \frac{N}{n} & g_{ks} = \frac{N_g}{N} \frac{n}{n_{s_g}} \\ \widehat{t}_{yr} = \sum_{g=1}^{G} N_g \overline{y}_{s_g} & \end{array} \tag{7.19}$$

où, comme on l'a déja observé, n_{s_g} est aléatoire.
Exercice.

1 Calculer les y^0_k (7.12) pour le modèle à moyenne de groupe avec plan SI.

2 En déduire que la variance approchée de (7.19) est :

$$\mathsf{var}_{\mathrm{app}}(\widehat{t}_{yr}) = N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_{g=1}^{G} (N_g - 1) S^2_{yU_g}$$

3 On veut estimer $AV(\widehat{t}_{yr})$. Pourquoi ne peut-on directement remplacer : $S^2_{yU_g}$ par $S^2_{ys_g}$?

4 Montrer que $\overline{e_{s_g}} = 0$. En déduire que $\overline{g_{ks}e_{ks}} = 0$.

5 Simplifiant, d'après le point précédent, l'expression générale de l'estimation de la variance approchée (), montrer que pour le plan SI, en supposant que $n(n_{s_g} - 1)/(n-1)n_{s_g}) \simeq 1$, l'approximation de la variance de l'estimateur du total est :

$$\widehat{V}_{\text{SI}}(\widehat{t}_{ypos}) = (1-f)\sum_{g=1}^{G} N_g^2 \frac{S^2_{y,s_g}}{n_{s_g}}$$

7.5.3 Modèle à ratio de groupe

Comme dans le modèle à moyenne de groupe, la population U est partitionnée en G sous-populations : $U_1, \cdots, U_G$, et le plan de sondage est défini sur U. Le modèle à ratio de groupe est :

$$\begin{aligned} E_\xi(y_k) &= \beta_g x_k \\ V_\xi(y_k) &= \sigma^2_{0g} x_k \end{aligned} \qquad k \in U_g, \;\; g = 1, \cdots, G \tag{7.20}$$

Ce modèle est utile si les ratios β_g sont très différents d'un groupe à l'autre. On introduit les variables auxiliaires :

$$x_{gk} = \begin{cases} x_k & \text{si} \quad k \in U_g \\ 0 & \text{si} \quad k \notin U_g \end{cases}$$

On obtient :

$$\mathbf{T} = \text{diag}\left(\sum_{U_1} \frac{x_k}{\sigma^2_{01}}, \cdots, \sum_{U_G} \frac{x_k}{\sigma^2_{0G}}\right) \mathbf{t} = \left[\sum_{U_1} \frac{y_k}{\sigma^2_{01}}, \cdots, \sum_{U_G} \frac{y_k}{\sigma^2_{0G}}\right]'$$

$$\mathbf{B} = \left[\frac{\sum_{s_1} \check{y}_k}{\sum_{s_1} \check{x}_k}, \cdots, \frac{\sum_{s_G} \check{y}_k}{\sum_{s_G} \check{x}_k}\right]$$

On a également :

$$g_{ks} = \frac{\sum_{U_g} x_k}{\sum_{s_g} \check{x}_k} = \frac{t_{xU_g}}{\widehat{t}_{xU_g}}, \quad k \in U_g.$$

On déduit immédiatement $\widehat{\mathbf{B}}$ et l'estimateur du total par régression :

$$\widehat{t}_{yr} = \sum_{g=1}^{G} \frac{t_{xU_g}}{\widehat{t}_{xU_g}} \sum_{s_g} \check{y}_k = \sum_{g=1}^{G} t_{xU_g} \frac{\widehat{t}_{yU_g}}{\widehat{t}_{xU_g}} \tag{7.21}$$

On appelle cet estimateur : **estimateur par ratio post stratifié**. On reconnaît dans (7.21) une somme d'estimateurs par ratio pour chaque groupe.

Cas particulier : plan SI. On tire l'échantillon dans U par un plan SI(N, n). Dans (7.21) les poids se simplifient. On obtient :

$$\widehat{t}_{yr} = \sum_{g=1}^{G} t_{xU_g} \frac{\sum_{s_g} y_k}{\sum_{s_g} x_k} \tag{7.22}$$

$$\widehat{\text{var}}_{SI}(\widehat{t}_{yr}) = (1-f)\sum_{g=1}^{G} (x_U/x_{s_g})^2 N_g^2 S^2_{e\,s_g}/n_{s_g}. \tag{7.23}$$

Dans l'expression de l'estimation de la variance, on a approché, $(n_{s_g} - 1)n/n_{s_g}n$ par 1. D'autre part,

$$S^2_{e\,s_g} = \frac{1}{n_{s_g} - 1} \sum\nolimits_{s_g} (y_k - \widehat{\mathbf{B}}_g x_k)^2$$

et $e_{k\,s_g} = y_k - \widehat{\mathbf{B}}_g x_k, \; = \sum\nolimits_{s_g} y_k / \sum\nolimits_{s_g} x_k.$