T. D. n° 7

Estimation par différence et par régression

Corrigé

Exercice 1. *Stratification et estimateur par la différence.* **D'après l'examen de Janvier 2006, M2-Statistique.**

1. L'estimateur est sans biais. En effet, puisque nous avons :

$$\widehat{\overline{Y}}_\pi = \sum_{h=1}^{H} \frac{N_h}{N} \widehat{\overline{Y}}_h,$$

où $\widehat{\overline{Y}}_h$ désigne la moyenne simple des y_k dans l'échantillon de la strate h,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\overline{Y}}_D\right) &= \mathbb{E}\left(\widehat{\overline{Y}}_\pi + \overline{X} - \widehat{\overline{X}}_\pi\right) \\
&= \mathbb{E}\left(\widehat{\overline{Y}}_\pi\right) + \overline{X} - \mathbb{E}\left(\widehat{\overline{X}}_\pi\right) \\
&= \overline{Y} + \overline{X} - \overline{X} \\
&= \overline{Y}.
\end{aligned}$$

2. Nous posons :

$$z_k = y_k - x_k.$$

Nous avons :

$$\widehat{\overline{Y}}_D = \overline{X} + \widehat{\overline{Z}}_\pi.$$

Donc nous pouvons calculer la variance de l'estimateur en se servant de cette décomposition :

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\overline{Y}}_D\right) &= \mathrm{Var}\left(\widehat{\overline{Z}}_\pi\right) \\
&= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{zh}^2}{n_h},
\end{aligned}$$

où S_{zh}^2 se définit par :

$$\begin{aligned}
S_{zh}^2 &= \frac{1}{N_h - 1} \sum_{k \in U_h} \left(z_k - \widehat{\overline{Z}}_h\right)^2 \\
&= S_{yh}^2 + S_{xh}^2 - 2S_{xyh},
\end{aligned}$$

et

$$S_{xyh} = \frac{1}{N_h - 1} \sum_{k \in U_h} \left(x_k - \widehat{\overline{X}}_h\right)\left(y_k - \widehat{\overline{Y}}_h\right).$$

3. En posant $z_k = y_k - x_k$, le problème revient à minimiser $\text{Var}\left(\widehat{\overline{Z}}_\pi\right)$ sous la seule contrainte de taille fixe, qui s'écrit ici $\sum_{h=1}^{H} n_h = n$. En effet, le coût unitaire est le même dans toutes les strates, ce qui donne :

$$n_h = \frac{N_h S_{zh}}{\sum_{l=1}^{H} N_l S_{zl}} n.$$

En pratique, nous estimons a priori les S_{zh}, nous arrondissons n_h à l'entier le plus proche, après avoir fixé n en fonction du budget global dont nous disposons. Il peut arriver que nous obtenons $n_h > N_h$ pour certains h : dans ce cas, nous imposons $n_h = N_h$ et nous reprenons l'ensemble du calcul avec les strates restantes.

4. Puisque la variance de l'estimateur s'écrit :

$$\text{Var}\left(\widehat{\overline{Y}}_\pi\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{yh}^2}{n_h},$$

et que les deux estimateurs sont sans biais, $\widehat{\overline{Y}}_D$ est indiscutablement préférable à $\widehat{\overline{Y}}_\pi$ lorsque, pour tout h :

$$S_{yh}^2 > S_{zh}^2,$$

soit, pour tout h :

$$\frac{S_{xyh}}{S_{xh}^2} > \frac{1}{2}.$$

Cette condition revient à obtenir une droite de régression de y sur x qui, dans chaque strate, ait une pente supérieure à 1/2. C'est en particulier le cas si nous posons $y = x$ (pente égale à 1) : ce résultat est naturel, car alors :

$$\widehat{\overline{X}}_D = \overline{X}$$

quel que soit l'échantillon tiré. Nous disons que l'estimateur $\widehat{\overline{Y}}_D$ est « calé » sur $\overline{X}$.

Exercice 2. Comparaison de plusieurs estimateurs.

1.(i) **Estimation par la moyenne à probabilités égales avec remise.** L'estimateur de la moyenne $\overline{X}$ est égal à :

$$\overline{x}_{SASPEAR} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

et sa variance est égale à :

$$\text{Var}\left(\overline{x}_{SASPEAR}\right) = \frac{\sigma_{pop}^2}{n}.$$

(ii) **Estimation par la moyenne à probabilités égales sans remise.** L'estimateur de la moyenne $\overline{X}$ est égal à :

$$\overline{x}_{SASPESR} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et sa variance est égale à :

$$\text{Var}\left[\overline{x}_{SASPESR}\right] = \frac{N-n}{N-1}\frac{\sigma^2_{pop}}{n}.$$

(iii) **Estimation par le quotient.**
L'estimateur de la moyenne $\overline{X}$ est égal à :

$$\overline{x}_Q = \overline{Y}\frac{\overline{x}}{\overline{y}}.$$

Cet estimateur est biaisé et par conséquent ce n'est pas la variance qu'il faut étudier mais l'erreur quadratique de cet estimateur.

2. Construisons l'estimateur par la différence $\overline{x}_D$:

$$\overline{x}_D = \overline{Y} + \overline{x}_{SAS} - \overline{y}_{SAS},$$

où $\overline{Y}$ dénote la moyenne des y pour la population entière.

Justification de l'estimateur. Il y a deux façons de justifier intuitivement cet estimateur.

Première justification. Nous pouvons décomposer $\overline{X}$ en deux parties, l'une connue, l'autre pas :

$$\overline{X} = \overline{Y} + \left(\overline{X} - \overline{Y}\right).$$

La première partie, $\overline{Y}$, est connue, et nous n'avons pas besoin de l'estimer. La deuxième partie, $\left(\overline{X} - \overline{Y}\right)$, la différence entre la moyenne des x et celle des y, n'est pas connue et doit être estimée. Nous l'estimons, naturellement, par la différence entre les deux moyennes échantillonnées, $\overline{x} - \overline{y}$. C'est ce qui donne l'estimateur par la différence.

Deuxième justification. L'estimateur naturel de $\overline{X}$ est $\overline{x}$ et a priori c'est l'estimateur privilégié. Dans l'estimateur par la différence, écrit comme :

$$\overline{x}_D = \overline{x} + \left(\overline{Y} - \overline{y}\right),$$

l'ajout du terme $\left(\overline{Y} - \overline{y}\right)$ peut s'interpréter comme un ajustement à l'estimateur $\overline{x}$. Grâce à notre information sur la variable y, nous pouvons deviner si, en l'occurrence, l'estimateur $\overline{x}$ a surestimé ou sous-estimé la moyenne $\overline{X}$.

3.(i) Calculons l'espérance de l'estimateur $\overline{x}_D$:

$$\begin{aligned}
\mathbb{E}\left(\overline{x}_D\right) &= \mathbb{E}\left(\overline{x}_{SAS} + \overline{Y} - \overline{y}_{SAS}\right) \\
&= \mathbb{E}(\overline{x}_{SAS}) + \overline{Y} - \mathbb{E}\left(\overline{y}_{SAS}\right) \\
&= \overline{X} + \overline{Y} - \overline{Y} \\
&= \overline{X}.
\end{aligned}$$

Nous en concluons que l'estimateur de la différence $\overline{x}_D$ est un estimateur sans biais de $\overline{X}$.

(ii) Calculons la variance de l'estimateur $\overline{x}_D$:

$$\begin{aligned}\text{Var}\left(\overline{x}_D\right) &= \text{Var}\left(\overline{x}_{SAS} + \overline{Y} - \overline{y}_{SAS}\right) \\ &= \text{Var}\left(\overline{x}_{SAS}\right) - 2\text{Cov}(\overline{x}_{SAS}, \overline{y}_{SAS}) + \text{Var}\left(\overline{y}_{SAS}\right).\end{aligned}$$

À partir de cette égalité, il est d'usage de se demander comment nous devons tirer l'échantillon : à probabilités égales avec ou sans remise.

4.(i) **Estimation par la moyenne à probabilités égales avec remise.**

$$22072,63.$$

(ii) **Estimation par la moyenne à probabilités égales sans remise.**

$$22072,63.$$

(iii) **Estimation par le quotient.**

$$32039,66 \times \left(\frac{22072,63}{21585,26}\right) = 32763,08.$$

(iv) **Estimateur par la différence.**

$$32039,66 + (22072,63 - 21585,26) = 32527,03.$$

Nous notons une grande variabilité dans les estimations. Pour choisir la meilleure estimation et par conséquent le meilleur estimateur, il faut savoir quelle est la plus petite variance associée à chaque estimateur. Pour cela, nous allons calculer les différents intervalles de confiance à 95%.

(i) **Intervalle de confiance pour la moyenne $\overline{X}$ (PEAR).**

$$\begin{aligned}22072,63 &\pm 1,96 \times \sqrt{\frac{4\,131\,789\,466}{35}} \\ 22072,63 &\pm 1,96 \times 10865,13 \\ 22072,63 &\pm 21295,67\end{aligned}$$

où $1,96$ est le quantile de la loi normale centrée réduite à 95%. Or nous faisons une large approximation en utilisant la loi normale centrée réduite. Il serait conseillé d'utiliser plutôt un quantile de Student à 95%.

(ii) **Intervalle de confiance pour la moyenne $\overline{X}$ (PESR).**

$$\begin{aligned}22072,63 &\pm 1,96 \times \sqrt{\left(1 - \frac{35}{180}\right)\frac{4\,131\,789\,466}{35}} \\ 22072,63 &\pm 1,96 \times 9751,76 \\ 22072,63 &\pm 19113,44\end{aligned}$$

où $1,96$ est le quantile de la loi normale centrée réduite à 95%. Or nous faisons une large approximation en utilisant une loi normale centrée réduite. Il serait conseillé d'utiliser plutôt un quantile de Student à 95%.

(iii) **Intervalle de confiance pour la moyenne $\overline{X}$ (par le quotient).** Estimons maintenant la variance de $\overline{x}_Q$:

$$\begin{aligned}\widehat{\text{Var}\,(\overline{x}_Q)} &= (1-f)\frac{s_x^2 - 2\widehat{R}s_{xy} + \widehat{R}^2 s_y^2}{n} \\ &= \left(1 - \frac{35}{180}\right)\frac{4\,131\,789\,466 - 2,04(4\,059\,448\,772) + (1,02)^2 3\,989\,656\,072}{35} \\ &= 32\,918,77.\end{aligned}$$

D'où l'intervalle de confiance pour la moyenne $\overline{X}$ est égal à :

$$\begin{aligned}32763,08 &\pm 1,96 \times 181,43 \\ 32763,08 &\pm 355,61.\end{aligned}$$

(iv) **Intervalle de confiance pour la moyenne $\overline{X}$ (par la différence).** Estimons maintenant la variance de $\overline{x}_D$:

$$\begin{aligned}\widehat{\text{Var}\,(\overline{x}_D)} &= (1-f)\frac{s_x^2 - 2s_{xy} + s_y^2}{n} \\ &= \left(1 - \frac{35}{180}\right)\frac{4\,131\,789\,466 - 2(4\,059\,448\,772) + 3\,989\,656\,072}{35} \\ &= 58\,644,31.\end{aligned}$$

D'où l'intervalle de confiance pour la moyenne $\overline{X}$ est égal à :

$$\begin{aligned}32527,03 &\pm 1,96 \times 242,17 \\ 32527,03 &\pm 474,64.\end{aligned}$$

Conclusion : Les deux derniers intervalles de confiance pour la moyenne $\overline{X}$ sont les plus intéressants pour nous car les deux écart-types sont très petits. D'autre part, il faut noter qu'il y a peu de différence entre l'estimateur par le quotient et l'estimateur par la différence.

5. Nous constatons que les deux estimateurs qui utilisent la variable auxiliaire sont très nettement meilleurs. Les estimations elles-mêmes sont très différentes : de l'ordre de 22 000 pour l'estimateur par la moyenne avec ou sans remise ; et de l'ordre de 32 000 pour les deux autres estimateurs. S'il fallait décider lequel des estimateurs à employer, nous dirions 32 000 plutôt que 22 000. Mais il n'est pas question, en pratique, de calculer plusieurs estimateurs et puis leur écart-type. Il faudrait pouvoir faire un choix d'avance. Est-ce que nous aurions pu prévoir la supériorité des deux derniers estimateurs, et s'en tenir à l'un ou l'autre de ces deux, sans même considérer le premier ? Dans le cas présent, nous aurions pu le prévoir. En général, les estimateurs qui font appel à une variable auxiliaire sont avantageux dans la mesure où la variable auxiliaire est corrélée positivement avec la variable d'intérêt. Il est évident que le nombre d'habitants en 1996 est corrélé avec le nombre d'habitants en 2001. C'est donc une information pertinente, et les estimateurs par la différence et par le quotient en tirent profit. La leçon importante qui se dégage est « si la variable auxiliaire est fortement liée à la variable d'intérêt, il vaut mieux utiliser l'estimateur par la différence ou l'estimateur par le quotient ».