Estimateur par régression rapide

Laurent BERREBI
Insee

I - INTRODUCTION

Cette note s'inscrit dans le cadre d'une étude méthodologique beaucoup plus générale, sur les enquêtes investissement et production dans l'industrie. Cette étude, dont l'objectif principal est d'améliorer la qualité des enquêtes, comportera trois étapes :

- analyse de la variance, proposition d'estimateurs et de méthodes d'échantillonnage plus efficaces, permettant d'améliorer la précision de l'estimation ;

- traitement des non-réponses ;

- utilisation des données de panel dans le but de modéliser les erreurs de prévision.

On s'intéresse ici au cas de l'enquête investissement, enquête quantitative dont la seule variable d'intérêt est la prévision de l'investissement.

Dans un premier temps, on fera l'analyse théorique d'estimateurs qui utilisent au mieux l'information disponible :ils seront comparés à l'estimateur actuellement utilisé, l'estimateur du ratio.

Dans un deuxième temps, on vérifiera en pratique les résultats théoriques obtenus et on proposera des méthodes d'estimation et d'échantillonnage plus efficaces.

II - ESTIMATEURS RESPECTANT UNE INFORMATION AUXILIAIRE BIAIS, VARIANCE, OPTIMALITE.

1) Introduction

Dans une population finie P, on se propose d'estimer le total $T_y = \sum_{i \in P} Y_i$ d'une variable Y à l'aide d'estimateurs linéaires, au vu d'un échantillon s aléatoire, obtenu par tirage au sort dans une loi connue. On note π_i la probabilité que l'individu i appartienne à l'échantillon, π_{ij} la probabilité que les individus i, j appartiennent à l'échantillon et ε_i la variable aléatoire qui vaut 1 si $i \in s$ et 0 sinon.

Un estimateur linéaire en y de T_y est une variable aléatoire de la forme : $\hat{Y}_w = \sum_{i \in s} w_i(s)\, Y_i$ (1)

$w_i(s)$ sont des poids aléatoires qui valent o si $i \notin s$

L'estimateur étant linéaire en y, les poids $w_i(s)$ sont indépendants des Y_i. Par contre, ils peuvent dépendre d'autres variables également mesurées dans le sondage. Toute l'information sera donc contenue dans le système de poids $w_i(s)$. On pourra alors estimer tout total T_X d'une variable X par :

$$\hat{X}_w = \sum_{i \in s} w_i(s)\, X_i$$

Si à chaque poids $w_i(s)$ est imposée l'indépendance vis à vis des autres individus sélectionnés dans l'échantillon, $w_i(s)$ peut s'écrire :

$$w_i(s) = w_i\, \varepsilon_i(s)$$

où $\varepsilon_i(s) = \begin{cases} 1 \text{ si } i \in s \\ 0 \text{ sinon} \end{cases}$

w_i est un coefficient qui n'est pas aléatoire

Dans ce cas, le seul estimateur sans biais est l'estimateur de Horvitz-Thompson encore appelé estimateur des valeurs dilatées :

$$\hat{Y}_{HT} = \sum_{i \in s} Y_i / \pi_i$$

Le poids ainsi attribué à chaque individu i, faisant partie de s, est d'autant plus grand que la probabilité de le tirer est petite.

La variance de cet estimateur est égale à :

$$V(\hat{Y}_{HT}) = \frac{1}{2} \sum_{i \neq j \in P} \sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

lorsque le plan de sondage est de taille fixe n.

Sous cette forme, il est évident qu'il existe un sondage de Horvitz-Thompson optimal, c'est-à-dire de variance minimale. En effet, si nous choisissons les probabilités d'inclusion π_i proportionnelles aux Y_i, la variance est nulle. Un tel sondage n'est évidemment pas utilisable, puisqu'il suppose connues les valeurs Y_i. Cependant, il est souvent possible de connaître une variable X presque proportionnelle à Y ; choisir les π_i homothétiques aux X_i permet alors d'approcher le sondage optimal.

Le choix des poids sera donc déterminé dans tous les cas en fonction des informations connues. Mais comment tenir compte de façon rationnelle et efficace de l'information dont on dispose ?

2) Utilisation d'une information auxiliaire.

On suppose connu à priori le total sur l'ensemble de la population P, de k variables également mesurées dans le sondage. Le vecteur $\underline{T}_x$ de ces totaux constitue ce qu'il convient d'appeler l'information auxiliaire. Pour utiliser de façon rationnelle cette information, on calibre l'estimateur (1), c'est-à-dire que l'on cherche des poids $w_i(s)$ pour lesquels les estimateurs $\underline{\hat{X}}_w$ estiment exactement $\underline{T}_x$:

$$(2) \qquad \underline{\hat{X}}_w = \sum_{i \in P} w_i(s)\, \underline{X}_i = \underline{T}_X \qquad \forall s$$

En particulier, cette condition signifie que la variance est nulle pour chacune des composantes de $\underline{\hat{X}}_w$:

$$V(\hat{X}_w) = E\left(\underline{\hat{X}}_w - \underline{T}_X\right)^2 = 0$$

exemples :

. L'estimateur du ratio s'utilise quand le total T_x d'une variable X mesurée dans l'enquête est connu .Il vérifie :

$$\hat{Y}_r = T_X\, \hat{Y}_{HT} / \hat{X}_{HT} = \sum_{i \in P} w_i(s)\, Y_i$$

où $$w_i(s) = \begin{cases} T_X \Big/ \left(\pi_i \sum_{j \in S} X_j / \pi_j\right) & \text{si } i \in S \\ 0 & \text{sinon} \end{cases}$$

Ce système de poids W permet d'estimer exactement T_X . En effet :

$$\hat{X}_w = \sum_{i \in S} w_i X_i = T_X \qquad \forall S$$

et

$$V(\hat{X}_w) = 0$$

. Dans le cas d'un échantillon équiprobable sans remise, l'estimateur de Horvitz-Thompson s'utilise lorsque le nombre total d'individus N est connu; le système de poids qu'il définit permet d'estimer exactement le nombre total d'individus dans l'échantillon. En effet, si X_i désigne la variable 1 constante :

$$\hat{X}_w = \sum_{i \in P} w_i(s)\, X_i = \sum_{i \in s} w_i(s) = N = T_X$$

puisque $w_i(s) = \begin{cases} w_i = N/n & \text{si } i \in s \\ 0 & \text{si } i \notin s \end{cases}$

En résumé, si le total sur l'ensemble de la population P de k variables est supposé connu, les poids $w_i(s)$ doivent vérifier l'ensemble des contraintes (2) et l'estimateur $\hat{Y}_w$ doit respecter autant que faire se peut l'absence de biais. Sachant que le système de poids $1/\pi_i$ vérifie déjà l'absence de biais, on voudrait, que les poids $w_i(s)$ soient le plus proche possible des poids $1/\pi_i$. $\varepsilon_i(s)$. L'idée consiste à se donner un système de nombres positifs a_i et une distance entre deux systèmes de poids $\underline{w}_1$ et $\underline{w}_2$ définie par :

$$D(\underline{w}_1, \underline{w}_2) = \sum_{i \in P} a_i\, E\left[\left(w_i^1(s) - w_i^2(s)\right)^2\right]$$

où E désigne l'espérance par rapport à la loi de probabilité de sélectionner l'échantillon s, soit p(s). On cherche donc à minimiser :

$$D(\underline{w}, \underline{d}) = \sum_{i \in P} a_i\, E\left[\left(w_i(s) - \varepsilon_i/\pi_i\right)^2\right] = \sum p(s) \sum_{i \in P} a_i (w_i(s) -$$

sous les contraintes

$$\sum_{i \in s} w_i(s)\, X_i = T_X \quad \forall s$$

Il suffit de résoudre le programme suivant pour chaque s, c'est-à-dire au vu de l'échantillon :

$$\begin{cases} \text{Min} \sum_{i \in s} a_i \left(w_i(s) - 1/\pi_i\right)^2 \\ \sum_{i \in s} w_i(s)\, X_i = T_X \end{cases}$$

3) Résolution : lien avec l'estimateur par la régression.

On montre (V. Deville 1988) que l'estimateur cherché :

$$\hat{Y}_w = \sum_{i \in \Sigma} w_i(s) Y_i$$

où les w_i sont solution du programme précédent, est "l'estimateur par régression" :

$$(3) \qquad \hat{Y}_w = \underline{w}'(s)\, Y = \hat{Y}_{HT} + \left(\underline{T}_X - \underline{\hat{X}}_{HT}\right)' \hat{b}_S$$

où

$$\hat{b}_S = \left(X' A^{-1} X\right)^{-1} X' A^{-1} Y_S$$

X, est la matrice des observations Xi

A la matrice diagonale indicée par les éléments de s, d'éléments a_i

Y_s, le vecteur indicé par s dont les éléments valent Y_i si $i \in S$ et 0 sinon.

$\hat{B}_s$ est un estimateur, au vu de l'échantillon s, du coefficient de régression pondérée b de Y sur X dans la population totale, chaque individu étant affecté du poids π_i/a_i. En effet, $X' A^{-1} X = \sum_{i \in s} X_i X_i'/a_i$ est un estimateur sans biais de $\sum_{i \in \Sigma} \pi_i X_i X_i'/a_i$, de même que $X' A^{-1} Y_S = \sum_{i \in s} X_i Y_i / a_i$ estime sans biais $\sum_{i \in \Sigma} \pi_i X_i Y_i / a_i$.

Donc, $\hat{B}_s$ est un estimateur asymptotiquement sans biais de

$$b = \left(\sum_{i \in \Sigma} \pi_i X_i X_i'/a_i\right)^{-1} \left(\sum_{i \in \Sigma} \pi_i X_i Y_i / a_i\right)$$

exemple :

Supposons qu'il n'y ait qu'une variable auxiliaire positive X. Dans ces conditions :

$$\hat{Y}_w = \sum_S Y_i/\pi_i + \left(T_X - \sum_S X_i/\pi_i\right) \frac{\sum_S X_i Y_i/a_i}{\sum_S X_i^2/a_i}$$

Si $a_i = \pi_i X_i$ (en supposant $X_i > 0$) on obtient l'estimateur par le ratio :

$$\hat{Y}_r = T_X \frac{\hat{Y}_{HT}}{\hat{X}_{HT}}$$

4) Propriétés de l'estimateur par régression.

$\hat{B}_S$ est un estimateur asymptotiquement sans biais de b. De plus, $\hat{\bar{Y}}_{HT}$ et $\hat{\bar{X}}_{HT}$ étant des estimateurs sans biais respectivement de T_y et $\underline{T}_x$, $\hat{\bar{Y}}_w$ est asymptotiquement sans biais, selon (3).

On peut réécrire (3) sous la forme :

$$\hat{\bar{Y}}_w = \underline{T}_x' b + (\underline{T}_x - \hat{\underline{X}}_{HT})' (\hat{b}_S - b) + \sum_{i \in S} U_i/\pi_i$$
$$= \underline{T}_x' b + (\underline{T}_x - \hat{\underline{X}}_{HT})' (\hat{b}_S - b) + \hat{U}_{HT}$$

où $U_i = Y_i - X_i b$ est le résidu, pour l'individu i, de la régression dans la population totale.

Le premier terme est non aléatoire. Si l'échantillon est de grande taille, le second terme est une variable aléatoire d'un ordre de grandeur très petit par rapport au troisième. La variance de $\bar{y}_w$ sera donc asymptotiquement donnée par la variance du troisième terme :

$$Vas(\hat{\bar{Y}}_w) = V(\hat{U}_{HT}) = \sum_{i,j \in P} c_{ij} U_i U_j$$

où

$$c_{ij} = E\left((w_i(S) - 1)(w_j(S) - 1)\right)$$

A ce stade, la question du choix des variables X_i les plus pertinentes se pose : quelles sont les variables, parmi toutes celles que l'on peut connaître, qui constitueront l'information auxiliaire ?

L'objectif est de minimiser la variance de l'estimateur : une façon d'y aboutir serait de diminuer les résidus. La corrélation des variables explicatives potentielles avec la variable d'intérêt Y paraît donc déterminante pour le choix de X_i : les variables, expliquant le mieux y, donc les plus corrélées à y, seront retenues.

5) Application à un tirage aléatoire simple sans remise.

. Supposons que le plan de sondage soit équiprobable sans remise de taille fixe n. Dans ce cas $\pi_i = n/N$.

$$\hat{U}_{HT} = \frac{N}{n} \sum_{i \in S} U_i$$

$$Vas(\hat{\bar{Y}}_w) = V(\hat{U}_{HT}) = \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{i \in P} (U_i - \bar{U})^2$$

$$\leq \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{i \in P} U_i^2$$

$$Vas(\hat{Y}_w) = V(\hat{U}_{HT}) \leq \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{i \in E} (Y_i - bX_i)^2$$

où $$b = \left(\sum_{i \in E} X_i X_i' / a_i \right)^{-1} \left(\sum_{i \in E} X_i Y_i / a_i \right) \qquad (4)$$

dans le cas de l'estimateur du ratio, $\sum_{i \in E} U_i = \sum_{i \in E} \left(Y_i - X_i \frac{\sum Y_i}{\sum X_i} \right)$

Donc : $$Vas(\hat{Y}_r) = V(\hat{U}_{HT}) = \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{i \in E} U_i^2$$

$$= \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{i \in E} (Y_i - b_r X_i$$

La quantité $\sum_{i \in E} (Y_i - bX_i)^2$ est minimale, quand b est egal à b_0, l'estimateur des moindres carrés ordinaires dans le modèle $Y_i = b\,X\,i + u_i$ (Il est obtenu en faisant $a_i = 1$ $\forall i$ dans l'expression (4).

Ainsi $$\sum_{i \in E} (Y_i - b_0 X_i)^2 \leq \sum_{i \in E} (Y_i - b_r X_i)^2$$

où $b_r = \frac{\sum_{i \in E} Y_i}{\sum_{i \in E} X_i}$ = b dans le cas où ai = Xi $\forall i$

(estimateur par le ratio).

Donc : $$V(\hat{Y}_0) \leq \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{i \in E} (Y_i - b_0 X_i)^2$$

$$\leq \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{i \in E} (Y_i - b_r X_i)^2$$

Par conséquent, si l'échantillon est de grande taille:

$$V(\hat{Y}_0) \leq V(\hat{Y}_r)$$

où $$\hat{Y}_0 = \frac{N}{n} \sum_{i \in S} Y_i + \left(T_x - \frac{N}{n} \sum_{i \in S} x_i \right) \frac{\sum_{i \in S} x_i Y_i}{\sum_{i \in S} x_i^2}$$

(estimateur par régression simple),

$$\hat{Y}_r = \frac{N}{n} \sum_{i \in S} Y_i + \left(T_x - \frac{N}{n} \sum_{i \in S} x_i \right) \frac{\sum_{i \in S} Y_i}{\sum_{i \in S} x_i}$$

(estimateur par le ratio)

dans le cas k = 1.

En pratique au vu d'un échantillon s, on estimera les variances par :

$$\hat{V}_S(\hat{Y}_0) = \frac{N(N-n)}{n(n-1)} \sum_{i \in S} \left(Y_i - \hat{b}_0^S x_i - \overline{U}_i^S \right)$$

$$\hat{V}_S(\hat{Y}_r) = \frac{N(N-n)}{n(n-1)} \sum_{i \in S} \left(Y_i - \hat{b}_r^S x_i \right)^2$$

Conclusion : l'estimateur par le ratio $\hat{Y}_r$ n'est pas l'estimateur optimal, l'estimateur par régression simple $\hat{Y}_0$ donnant toujours une variance $V(\hat{Y}_0)$ au plus égale, sinon inférieure à $V(\hat{Y}_r)$.

On s'est ainsi attaché dans cette partie à montrer théoriquement que l'estimateur par le ratio n'est pas l'estimateur optimal du total T_y et que les estimateurs par régression simple sont plus efficaces.

Il reste à vérifier si cette propriété est observable en pratique.

III - APPLICATION A L'ESTIMATION DU TAUX D'ACCROISSEMENT DE L'INVESTISSEMENT

1) Introduction

L'objectif est ici d'appliquer les estimateurs vus précédemment dans le cas où la variable d'intérêt Y est l'investissement de l'année n en cours : on cherche à estimer le taux d'accroissement d'investissement entre l'année n-1 et l'année n, au vu d'un échantillon s, avec le maximum de précision possible.

Il faut bien préciser que les échantillons sont stratifiés, pour les besoins de l'enquête d'une part - nécessité de connaître le taux d'accroissement de l'investissement par secteur - et pour les besoins statistiques d'autre part - une stratification en strates homogènes permet de diminuer le nombre d'entreprises nécessaire à une estimation de bonne qualité. Or, la pratique d'échantillons stratifiés nécessite la connaissance des critères optimaux de stratification, puis le calcul du nombre optimal nh d'individus par strates, la taille de l'échantillon n étant fixée.

Pour l'usage de telles méthodes, il est tout à fait essentiel de connaître au départ la population exhaustive afin de déterminer un plan de sondage optimal. Pour cela, j'ai pris l'EAE 85 et l'EAE 86, les deux dernières enquêtes annuelles d'entreprises dont les résultats sont définitifs. Il a fallu retirer toutes les entreprises sur lesquelles l'enquête est quasiment muette : il s'agit d'entreprises de petite taille, de moins de 20 salariés en général, auxquelles sont envoyés des questionnaires "minimum", qui ne renseignent sur rien, exceptés l'effectif et le numéro de siren. Le numéro de siren a permis de réunir les deux enquêtes et de constituer un fichier de données de panels qui comporte 23 800 entreprises.

Dans un premier temps, il est nécessaire de choisir les variables de stratification les plus pertinentes, pour chaque estimateur utilisé. Des échantillons seront ensuite tirés dans chaque strate de manière optimale, la taille de l'échantillon total n étant fixée. Enfin, la variance de chaque estimateur, développé ci-après, sera estimée au vu de l'échantillon tiré, dans le but de sélectionner un estimateur de meilleure précision que l'estimateur utilisé actuellement.

2) Estimateurs utilisés.

On suppose que la population et l'échantillon sont subdivisés en H strates, respectivement $P_1, \ldots, P_H$ et $S_1, \ldots, S_H$ quatre estimateurs vont être utilisés afin d'estimer $Ty = \sum_{i \in P} Y_i$

. L'estimateur de Horvitz-Thompson noté (1)

$$(1) \qquad \hat{Y}_{HT} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in S_h} Y_i$$

de variance

$$V(\hat{Y}_{HT}) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h} S_h^2$$

où

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in P_h} (Y_i - \bar{Y}_h)^2, \quad \bar{Y}_h = \frac{1}{N_h} \sum_{i \in P_h} Y_i$$

et de variance estimée :

$$\hat{V}_S(\hat{Y}_{HT}) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h} s_h^2$$

où

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i \in S_h} (y_i - \bar{y}_h)^2 \text{ où } \bar{y}_h = \frac{1}{n_h} \sum_{i \in S_h} y_i$$

. L'estimateur par le ratio utilisé actuellement noté (2) :

$$(2) \qquad \hat{Y}_r = \sum_{h=1}^{H} \left[\left(\frac{\sum_{i \in S_h} Y_i}{\sum_{i \in S_h} X_i} \right) \sum_{i \in P_h} X_i \right]$$

où X_i représente l'investissement de l'année passée n-1. Sa variance, d'après II.4), est :

$$V(\hat{Y}_r) = \sum_{h=1}^{H} \left(\frac{\sum_{i \in S_h} Y_i}{\sum_{i \in S_h} X_i} \right) \times \sum_{i \in P_h} X_i$$

où

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in P_h} (U_i - \bar{U}_h)^2$$

$$U_i = Y_i - \frac{\sum_{i \in P_h} Y_i}{\sum_{i \in P_h} X_i} X_i$$

(résidu de la régression dans la population totale)

$$\bar{U}_h = \frac{1}{N_h} \sum_{i \in P_h} U_i$$

D'après II;4), si chaque strate de l'échantillon est de grande taille, $\sum_{i \in s_h} Y_i / \sum_{i \in s_h} X_i$ est un estimateur convergent de $\sum_{i \in P_h} Y_i / \sum_{i \in P_h} X_i$ et $V(\hat{Y}_r)$ peut être ainsi estimée au vu de l'échantillon par :

$$\hat{V}_S(\hat{Y}_r) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h} s_h^2$$

où

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (u_i - \bar{u}_h)^2$$

$$u_i = Y_i - \frac{\sum_{i \in s_h} Y_i}{\sum_{i \in s_h} X_i} X_i \quad \text{et} \quad \bar{u}_h = \frac{1}{n_h} \sum_{i \in s_h} u_i$$

. L'estimateur par régression simple, noté (3), en tenant compte de l'investissement de l'année passée, année n-1, en tant qu'information auxiliaire. On note :

$$\hat{Y}^h_{m.c.o.} = \hat{Y}^h_{HT} + \left(T^h_X - \hat{X}^h_{HT}\right) \frac{\sum_{i \in s_h} X_i Y_i}{\sum_{i \in s_h} X_i^2}$$

l'estimateur par régression simple de la masse totale d'investissement T_y^h de la state h.

La variance égale à

$$V(\hat{Y}^h_{m.c.o}) = \frac{N_h(N_h - n_h)}{n_h} S_h^2$$

où

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in P_h} (U_i - \bar{U}_h)^2, \quad U_i = Y_i - \frac{\sum_{i \in P_h} X_i Y_i}{\sum_{i \in P_h} X_i^2} \quad \forall i \in$$

peut être estimée, d'après II.4) si l'échantillon est de grande taille, par

$$\hat{V}_S(\hat{Y}^h_{m.c.c.}) = \frac{N_h(N_h - n_h)}{n_h} s_h^2$$

où,

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (u_i - \bar{u}_h)^2, \quad u_i = y_i - \frac{\sum_{i \in s_h} x_i y_i}{\sum_{i \in s_h} x_i^2} x_i$$

L'estimateur du total T_y sera donc :

$$\hat{Y}_{m.c.o.} = \sum_{h=1}^{H} \hat{Y}^h_{m.c.o}$$

de variance estimée

$$\hat{V}_s(\hat{Y}_{m.c.o}) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h} s_h^2 = \sum_{1}^{H} \hat{V}_s(\hat{Y}^h_{m.c.o})$$

. L'estimateur par régression simple (4), l'information auxiliaire étant constituée de l'investissement de l'année passée et du chiffre d'affaires de l'année passée.

$$\hat{Y}_2^h = \hat{Y}^h_{HT} + \left(\bar{X}^h - \hat{X}^h_{HT}\right)\left(X'_h X_h\right)^{-1} X'_h Y_h^s$$

de variance

$$V(\hat{Y}_2^h) = \frac{N_h(N_h - n_h)}{n_h} S_h^2$$

où

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in P_h} (U_i - \bar{U}_h)^2 \quad \text{avec } U_i = Y_i - X_h(X'_h X_h)^{-1} X_i$$

de variance estimée

$$\hat{V}_s(\hat{Y}_2^h) = \frac{N_h(N_h - n_h)}{n_h} s_h^2$$

L'estimateur du total T_y $\hat{Y}_2 = \sum_{1}^{H} \hat{Y}_2^h$ aura donc pour variance :

$$V(\hat{Y}_2) = \sum_{1}^{H} \frac{N_h(N_h - n_h)}{n_h} S_h^2$$

estimée par

$$\hat{V}_s(\hat{Y}_2) = \sum_{1}^{H} \frac{N_h(N_h - n_h)}{n_h} s_h^2$$

Pour estimer le taux d'accroissement t de l'investissement, il suffit de prendre l'estimateur $\hat{Y}$ de T_y que l'on divise par la masse totale d'investissement de l'année précédente M_o :

$$\hat{t} = \frac{\hat{Y}}{M_o}$$

de variance

$$V(\hat{t}) = \frac{V(\hat{Y})}{M_o^2}$$

estimée par

$$\hat{V}_S(\hat{F}) = \frac{\hat{V}_S(\hat{Y})}{M_0^2}$$

Les estimateurs (1), (2), (3), (4) seront les estimateurs utilisés par la suite. Avant d'aller plus loin, il peut être intéressant de donner une signification aux estimateurs en particulier aux estimateurs (1), (2), (3).

L'estimateur (1) peut être réécrit :

$$\hat{Y}_{HT} = \sum_{h=1}^{H} \left((N_h - n_h) \frac{\sum_{i \in s_h} Y_i}{n_h} + \sum_{i \in s_h} Y_i \right)$$

$\sum_{i \in s_h} Y_i$ représente la masse d'investissement des entreprises faisant partie de l'échantillon.

L'estimateur (1) consiste donc à attribuer aux $(N_h - n_h)$ entreprises appartenant à la strate h et ne faisant pas partie de l'échantillon, l'investissement moyen de la strate calculé à partir de l'échantillon, soit $\sum_{i \in s_h} Y_i / n_h$

L'estimateur (2) peut être réécrit de la façon suivante, dans chaque strate h :

$$\hat{Y}_{\pi}^{h} = \frac{N_h}{n_h} \sum_{i \in s_h} Y_i + \left(\sum_{i \in P_h} X_i - \frac{N_h}{n_h} \sum_{i \in s_h} X_i \right) \sum_{i \in s_h} \frac{Y_i}{X_i} \frac{X_i}{\sum_{i \in s_h} X_i}$$

$$= \frac{N_h}{n_h} \sum_{i \in s_h} Y_i + \left(\sum_{i \in P_h} X_i - \frac{N_h}{n_h} \sum_{i \in s_h} X_i \right) i_1^h$$

où i_1^h est le taux d'accroissement moyen de la strate h dans l'échantillon, pondéré par les poids $X_i / \sum_{i \in s_h} X_i$

Or, $\sum_{i \in P_h} X_i - (N_h/n_h) \sum_{i \in s_h} X_i$ représente la masse d'investissement de l'année (n-1) dont on n'a pas tenu compte en utilisant l'estimateur (1) : c'est en fait l'erreur commise l'année précédente sur la masse totale d'investissement de la strate h.

(2) revient donc à corriger (1) en affectant à cette erreur un taux d'accroissement moyen i_1^h.

De même, l'estimateur (3) peut s'écrire :

$$\hat{Y}_{m.c.c}^{h} = \frac{N_h}{n_h} \sum_{i \in s_h} Y_i + \left(\sum_{i \in P_h} X_i - \frac{N_h}{n_h} \sum_{i \in s_h} X_i \right) \sum_{i \in s_h} \frac{Y_i}{X_i} \frac{X_i^2}{\sum_{i \in s_h} X}$$

Il consiste donc à corriger (1) en affectant le taux d'accroissement pondéré de la strate h

$$\lambda_2^h = \sum_{i \in s_h} \left(\frac{Y_i}{X_i} \right) \times \frac{X_i^2}{\sum_{i \in s_h} X_i^2}$$

à l'erreur commise.

3) Stratification : critères optimaux et allocations optimales

A partir de l'estimateur choisi, il faut définir un plan de sondage optimal, d'une part en choisissant les critères optimaux de stratification, d'autre part en répartissant dans chaque strate, de façon optimale, le nombre n d'entreprises de l'échantillon.

Or, la stratification est d'autant plus efficace que les strates obtenues sont homogènes, c'est-à-dire que la variabilité de chaque strate est petite : c'est l'un des principaux résultats de la théorie de la stratification.

Mais de quelle variabilité s'agit-il ? Il faut bien se rendre compte que dans le cas de l'estimateur (1), la variance est déterminée à partir de la variance de la variable Y. Par contre, dans les cas (2), (3) et (4), la variance est calculée à partir de la variance de résidus. Par conséquent, le gain de précision, obtenu par stratification, sera d'autant plus important que les variables de stratification seront plus corrélées avec Y dans le cas (1) et avec les résidus U dans les cas (2), (3), (4).

Je me suis limité ici aux estimateurs (2) et (3), et à deux variables potentielles de stratification, mise à part la variable secteur qui s'impose tout naturellement : les effectifs à la fin de l'année n-1, et le chiffre d'affaires de l'année n-1. Les corrélations par secteur d'activité au niveau 15 entre les résidus et les variables précédemment définies ne permettent pas d'en choisir une plutôt que l'autre, comme le montrent les tableaux suivants.

Tableau 1 : corrélation entre résidus et effectif, chiffre d'affaires suivant l'estimateur utilisé.

		U_1 (résidus par l'estimateur (3)	U_2 (résidus par l'estimateur (2)
Industries agricoles et alimentaires	Eff	0,11	- 0,15
	CA	0,08	- 0,15
Industries des biens intermédiaires	Eff	- 0,2	- 0,14
	CA	- 0,34	- 0,23
Industries des biens d'équipement professionnels	Eff	0,15	0,28
	CA	- 0,01	- 0,11
Industries automobile et transport terrestre	Eff	0,49	0,16
	CA	0,21	- 0,16
Industries des biens de consommation courante et des biens d'équipement ménager	Eff	0,08	- 0,19
	CA	0,19	- 0,06

Pour cette raison, j'ai gardé comme variable de stratification l'effectif. Chaque strate est par conséquent définie par son secteur et sa tranche de taille.

Se pose aussi le problème du découpage optimal : pour un nombre de strates donné, quelles sont les valeurs limites de la variable de stratification déterminant les strates optimales ?

Dalenius a développé une méthode (W. Cochran 1977) : on ne s'étendra pas ici sur le sujet car la méthode appliquée à notre population a donné des résultats moyens, sans doute à cause du manque de corrélation entre les résidus et la variable de stratification, c'est-à-dire l'effectif.

Quant au nombre de strates à choisir, j'ai gardé aussi 3 tranches de taille par niveau 15. En effet, le gain de précision en stratifiant davantage est très minime dès que le nombre de tranches est supérieur ou égal à 3. Les tranches de taille retenues sont les suivantes : 0-100, 100-500, 500 et plus.

Il reste à définir l'allocation optimale par strate. La fonction de coût pour observer l'échantillon s de taille totale n et de taille n_h dans la strate h est :

$$C(s) = C_0 + \sum_{h=1}^{H} C n_h$$

C étant le coût du timbre et de l'enveloppe.

Si l'on dispose d'une somme C_1 pour réaliser l'enquête, on choisira la répartition optimale n_h, h = 1 H permettant d'obtenir la meilleure précision tout en satisfaisant la contrainte $C(s) < C_1$. Il faut donc résoudre le programme suivant :

$$\begin{cases} \underset{n_1 \ldots n_H}{\text{Min}} \sum_{1}^{H} \dfrac{N_h (N_h - n_h)}{n_h} S^2_{h,s} & \text{suivant les cas } S_h^2 = (S_h^y)^2 \text{ ou } (S_h^u)^2 \\ \text{sous } C(s) = C_0 + \sum_{1}^{H} C n_h = C_1 \text{ (contrainte saturée à l'optimum)} \end{cases}$$

La condition d'optimalité s'écrit : n_h/N_h proportionnel à $\sqrt{S_h^2} = S_h$

Dans ce cas, la variance de l'estimateur est égale à :

$$V_{opt} = \frac{\left(\sum_{1}^{H} N_h S_h\right)^2}{N} - \sum_{1}^{H} N_h S_h^2$$

En particulier, la notion de représentativité de l'échantillon est denuée de tout fondement : l'échantillon optimal (à taille n donnée) est en général non représentatif si par représentativité l'on sous-entend soit un taux de sondage constant par strate $\frac{n_h}{n} = \frac{N_h}{N}$ soit un taux de couverture constant par strate

$$\sum_{i \in S_h} Y_i \Big/ \sum_{i \in S} Y_i = \sum_{i \in P_h} Y_i \Big/ \sum_{i \in P} Y_i$$

4) Tests sur échantillons.

Dans les tableaux 2 (2.1, 2.2, 2.3, 2.4, 2.5, 2.6) sont inscrits les écarts-type de chaque estimateur, pour une taille d'échantillon n = 2 500, 4 000, 5 000, 6 000, 7 500.

Dans les tableaux 3, sont inscrits les écarts-type estimés des taux d'accroissement estimés, total et sectoriels, de l'investissement de l'année 1985 à 1986 à partir d'échantillons tirés dans une loi équiprobable sans remise.

Tableau 2.1 : Ecart-type du taux d'accroissement total estimé.

nbre d'entreprises de l'échantillon \ estimateur	1	2	3	4
2 500	2,7	2,0	1,6	1,6
4 000	1,9	1,3	1,1	1,1
5 000	1,6	1,0	0,9	0,9
6 000	1,3	0,9	0,8	0,7
7 500	1,1	0,7	0,6	0,6

Tableau 2.2 : Ecart-type du taux d'accroissement estimé dans le secteur agro alimentaire

n \ estimateur	1	2	3	4
2 500	7,3	5,6	4,8	4,6
4 000	4,9	3,8	3,2	3,1
5 000	4,0	3,2	2,6	2,5
6 000	3,4	2,7	2,2	2,1
7 500	2,6	1,8	1,7	1,6

Tableau 2.3 : Ecart-type du taux d'accroissement estimé dans le secteur des biens intermédiaires + énergie.

n \ estimateur	1	2	3	4
2 500	4,2	2,8	2,2	2,1
4 000	2,9	1,7	1,5	1,4
5 000	2,4	1,2	1,2	1,1
6 000	2,0	1,0	1,0	0,9
7 500	1,5	0,8	0,7	0,7

Tableau 2.4 : Ecart-type du taux d'accroissement estimé dans le secteur des biens d'équipement professionnels

n \ estimateur	1	2	3	4
2 500	4,9	3,8	3,1	3,0
4 000	3,5	2,7	2,2	2,1
5 000	2,9	2,3	1,8	1,7
6 000	2,5	1,9	1,5	1,4
7 500	2,0	1,3	1,1	1,0

Tableau 2.5 : Ecart-type du taux d'accroissement estimé dans le secteur du matériel de transport terrestre

n \ estimateur	1	2	3	4
2 500	7,8	6,3	4,6	4,7
4 000	5,1	4,2	3,2	3,0
5 000	4,3	3,6	2,7	2,6
6 000	3,8	3,2	2,4	2,2
7 500	3,2	2,4	1,9	1,8

Tableau 2.6 : Ecart-type du taux d'accroissement estimé dans le secteur des biens de consommation + biens d'équipement ménager.

n \ estimateur	1	2	3	4
2 500	7,8	5,7	4,9	4,7
4 000	5,3	4,0	3,4	3,3
5 000	4,5	3,3	2,8	2,7
6 000	4,0	2,9	2,4	2,3
7 500	3,3	2,1	2.0	1.9

Tableau 3.1 : Ecart-type estimé du taux d'accroissement global estimé

n / estimateur utilisé	(1)	(2)	(3)	(4)
4 000	2,4	1,5	1,2	1,1
5 000	1,9	1,2	1,0	0,9
6 000	1,3	0,9	0,8	0,8

Tableau 3.2 : Ecart-type estimé du taux d'accroissement estimé dans le secteur agro alimentaire (N_1 = 3 821)

n / estimateur utilisé	(1)	(2)	(3)	(4)
4 000 (n_1 = 673)	6,0	5,1	5,1	5
5 000 (n_1 = 842)	4,9	4,0	3,9	3,9
6 000 (n_1 = 1 008)	4,0	3,2	3,2	3,2

Tableau 3.3 : Ecart-type estimé du taux d'accroissement estimé dans le secteur des biens intermédiaires + énergie (N_2 = 6 375)

n \ estimateur utilisé	(1)	(2)	(3)	(4)
4 000 (n_2 = 1 285)	4,2	2,4	1,4	1,4
5 000 (n_2 = 1 604)	3,6	1,9	1,2	1,1
6 000 (n_2 = 1 916)	2,9	1,4	1,0	0,9

Tableau 3.4 : Ecart-type estimé du taux d'accroissement estimé dans le secteur des biens d'équipement professionnels (N_3 = 4 484)

n \ estimateur utilisé	(1)	(2)	(3)	(4)
4 000 (n_3 = 796)	3,4	2,9	2,9	2,2
5 000 (n_3 = 963)	2,7	2,3	2,3	1,8
6 000 (n_3 = 1 127)	2,1	1,8	1,8	1,5

Tableau 3.5 : Ecart-type estimé du taux d'accroissement estimé dans le secteur automobile (N_4 = 543)

n \ estimateur utilisé	(1)	(2)	(3)	(4)
4 000 (n_4 = 122)	2,8	2,1	2,0	1,8
5 000 (n_4 = 129)	2,6	2,2	1,9	1,8
6 000 (n_4 = 145)	2,5	2,2	1,8	1,5

Tableau 3.6 : Ecart-type estimé du taux d'accroissement estimé dans le secteur des biens de consommation + biens d'équipement ménager (N_5 = 8 438)

n \ estimateur utilisé	(1)	(2)	(3)	(4)
4 000 (n_5 = 1 924)	3,6	3,1	2,7	2,5
5 000 (n_5 = 1 427)	3,0	2,6	2,3	2,2
6 000 (n_5 = 1 696)	2,4	2,3	2,1	2,0

5) Commentaires et conclusion.

A la lecture des tableaux précédents, plusieurs commentaires peuvent être faits.

. Les tests confirment bien que les estimateurs (3) et (4) sont plus efficaces que l'estimateur par le ratio. Si on se réfère par exemple non pas aux écarts-types estimés mais aux valeurs des écarts-types des tableaux 2, on voit par exemple que pour un échantillon de 5 000 entreprises, la précision des estimations dans les cas (3) et (4) est au moins aussi bonne, sinon meilleure dans certains secteurs, que la précision fournie par l'estimateur du ratio pour un échantillon de 6 000 entreprises, soit mille entreprises de plus.

Par contre, les écarts-type des estimateurs (3) et (4) sont pratiquement les mêmes (une différence de 0,1 ou 0,2 en faveur de (4), selon les cas).

. Ici, seuls trois exemples d'échantillons ont été donnés (n = 4 000, 5 000, 6 000). D'autres échantillons ont été tirés afin de vérifier, en particulier, si l'intervalle de confiance $[\hat{Y} - 2\hat{\sigma}, \hat{Y} + 2\hat{\sigma}]$, $\hat{\sigma}$ désignant l'écart-type de l'estimateur contenait la valeur du taux d'accroissement calculé sur la population exhaustive, ce qui revient à tester si on peut accepter ou refuser l'hypothèse d'absence de biais des estimateurs utilisés. Il convient cependant d'être prudent sur la robustesse des estimateurs.

En effet, on peut observer dans certains secteurs une différence assez considérable entre l'écart-type et l'écart-type estimé, par exemple dans le secteur agro-alimentaire. Or, l'influence d'un point peut être très importante. Il suffit de regarder la distribution des résidus pour s'en convaincre : un résidu atypique d'une certaine entreprise dans une strate peut influencer, de manière non négligeable, l'écart-type estimé selon que l'entreprise fasse partie ou non de l'échantillon. De même, il peut être possible que la valeur du taux d'accroissement ne soit pas à l'intérieur de l'intervalle de confiance à cause d'une entreprise atypique.

Pour y remédier, on pourrait penser soit à utiliser le "bootstrap" ou le "jacknife" qui ont l'avantage de réduire un biais éventuel, soit à traiter les entreprises atypiques de manière différente après les avoir repérées : pour les repérer, l'option "influence" de la "Proc Reg" de SAS permettrait de connaître l'influence de chaque point sur les coefficients de régression et de déterminer ainsi les points atypiques.

Exemple : n = 5 000 entreprises

	t	$t_4 \pm 2\hat{\sigma}_2$	$t_3 \pm 2\hat{\sigma}_3$
Population totale	5,5	3,9 (± 2,4)	4,4 (± 2,0)
Secteur agro-alimentaire	- 2,2	- 3,7 (± 8,1)	- 4,0 (± 7,8)
Biens intermédiaires + énergie	6,0	4,2 (± 3,8	5,5 (± 2,4)
Biens d'équipement profes-nels	12,7	13,6 (± 4,7)	13,7 (± 4,7)
Matériel de transport	0,6	3,2 (± 4,4)	1,7 (± 3,8)
Biens de consommation + biens d'équipement ménager	3,7	0,2 (± 5,2)	0,6 (± 4,4)

. L'allocation de Neyman a mené à un tirage exhaustif dans les petites strates, c'est-à-dire pour les grosses entreprises (entreprises de plus de 500 salariés). Il est certain que certaines entreprises parmi les grosses ne répondront pas à l'enquête et que la variance dans ces strates sera non nulle.

. Il peut être envisageable de prendre en compte d'autres informations auxiliaires, comme d'autres variables retardées de l'investissement, par exemple l'investissement de l'année n-2. Mais, il faudra déjà construire un panel exhaustif, ce qui n'est pas simple, le panel étant constitué à partir du n° de siren (le nombre d'entreprises qui changent de n° de siren, qui se restructurent ou qui se créent n'est pas du tout négligeable sur deux ans). Ceci dit, pour tester des estimateurs, il suffit de tirer des échantillons dans une population que l'on peut considérer exhaustive.

. On peut se demander si l'allocation de Neymam est stable dans le temps. Pour cela, on pourra par exemple calculer les écarts-types et les nombres d'entreprises par strate sur les années 84-85, 85-86, 86-87, lisser en faisant la moyenne sur les trois fichiers:

$$\left(\sum_{t=1}^{3} N_h^t \sqrt{S_h^t} \right) / 3$$

et avoir une allocation de la forme :

$$n_h = n \frac{\sum_{t=1}^{3} N_h^t \sqrt{S_h^t}}{\sum_{t=1}^{3} \left(\sum_{h=1}^{H} N_h^t \sqrt{S_h^t} \right)}$$

. L'estimateur du ratio possède une propriété "économique" que les autres n'ont pas : c'est un estimateur d'agrégation. En effet, si le taux d'investissement par strate des entreprises dans l'échantillon est i_h, la masse d'investissement de la strate h est estimée par la quantité $M_n \times i_h$ où M_n est la masse totale d'investissement de la strate h de l'année précédente. Les autres estimateurs ne possèdent pas cette propriété.

. Dans le traitement actuel de l'enquête, les gros investisseurs qui sont des entreprises non extrapolables sont séparées des entreprises extrapolables. Il est tout à fait possible de réintégrer les gros investisseurs qui sont des entreprises de plus de 500 salariés, dans la population exhaustive de départ : l'échantillon étant exhaustif pour les strates de petite taille regroupant les grosses entreprises, le résultat ne changera pas.

. En estimant de plusieurs façons le taux d'investissement, la technique de recouvrement des intervalles de confiance pourrait être utilisée : le taux d'investissement cherché appartenant à l'intersection des intervalles de confiance.