

MAMADOU SALIOU DIALLO

# L'approche prédictive dans la théorie de l'échantillonnage

Essai présenté  
à la Faculté des études supérieures de l'Université Laval  
dans le cadre du programme de maîtrise en statistique  
pour l'obtention du grade de Maître ès sciences (M.Sc.)

Département de mathématiques et de statistique  
FACULTÉ DES SCIENCES ET DE GÉNIE  
UNIVERSITÉ LAVAL  
QUÉBEC

Avril 2006

©Mamadou Saliou Diallo, 2006

# Résumé

Ce travail porte sur l'échantillonnage selon le modèle linéaire général. La théorie est présentée aussi bien dans le cas des variables continues que dans celui des variables qualitatives (facteurs). Le théorème général de la prédiction (Royall 1976) donne le meilleur prédicteur linéaire non biaisé, meilleur dans le sens où parmi tous les estimateurs non biaisés sa variance est minimale. Les échantillons équilibrés, par rapport à nos variables auxiliaires, permettent de conserver la caractèrè non biaisé de nos estimateurs quand le modèle est "faux". Le modèle est dit "faux" quand il ne représente pas bien les données. Une simulation est effectuée pour un modèle possédant deux variables qualitatives avec un nombre assez grand de niveaux (30) et une seule observation par cellule. Les cellules sont formées par le croisement des niveaux des deux facteurs (tableau de contingence).

# Avant-propos

Avant tout j'aimerais remercier mon directeur de recherche, Monsieur Louis-Paul Rivest sans qui tout ce travail n'aurait été aussi agréable. Sa grande connaissance du sujet m'a permis d'avancer sereinement et efficacement. Je tiens aussi à louer M. Rivest et le Département de mathématiques et de statistique pour leur effort d'octroyer des bourses de recherches car sans cette aide financière, je ne pouvais pas me consacrer entièrement à mes études.

Papa et Maman je vous remercie du fond du coeur pour votre présence en dépit de la distance qui nous sépare mais aussi pour votre encouragement et l'éducation que vous m'avez inculqué qui me sert chaque jour que Dieu fait. Merci à mes frères et soeurs d'être mes premiers amis et confidents. Bien sur vous mes amis du Senegal, du Canada, de la France et ailleurs dans le monde, vous occupez une place spécial dans ma vie et grâce à vous chaque jour le soleil brille dans mon coeur, merci pour toutes ces belles journées.

# Table des matières

|  |           |
|--|-----------|
| Résumé   | ii        |
| Avant-Propos   | iii       |
| Table des matières   | v         |
| Liste des tableaux   | vi        |
| Table des figures  | vii       |
| Introduction   | 1         |
| <b>1 Modèles avec variables quantitatives</b>  | <b>3</b>  |
| 1.1 Introduction à l'approche par le modèle . . . . .  | 3         |
| 1.1.1 Notation . . . . .   | 4         |
| 1.1.2 Qu'est une approche par le modèle . . . . .  | 5         |
| 1.1.3 Un exemple de l'approche par le modèle . . . . .   | 5         |
| 1.1.4 Pourquoi une approche par le modèle . . . . .  | 9         |
| 1.2 L'approche prédictive sous le modèle linéaire général . . . . .                                | 9         |
| 1.2.1 Définitions . . . . .  | 10        |
| 1.2.2 Théorème général de la prédiction (Royall 1976b) . . . . .                                   | 11        |
| 1.2.3 Le meilleur prédicteur linéaire non-biaisé sous quelques modèles<br>simples . . . . .        | 14        |
| 1.3 Approche prédictive et échantillons équilibrés . . . . .                                       | 18        |
| 1.3.1 Robustesse et biais . . . . .  | 18        |
| 1.3.2 Définition d'un échantillon équilibré et de la stratégie de robustesse<br>au biais . . . . . | 21        |
| 1.3.3 Quelques estimateurs sous échantillons équilibrés . . . . .                                  | 23        |
| <b>2 Modèles avec variables qualitatives</b>   | <b>25</b> |
| 2.1 Introduction . . . . .   | 25        |
| 2.1.1 Exemple simple . . . . .   | 26        |
| 2.1.2 Facteur, niveau et effet . . . . .   | 28        |

|          |   |           |
|----------|---|-----------|
| 2.2      | Inverse généralisé . . . . .  | 29        |
| 2.3      | Estimation d'une combinaison linéaire des $Y_i$ . . . . .   | 32        |
| 2.3.1    | Cas général . . . . .   | 32        |
| 2.3.2    | Estimation du total dans un modèle unifactoriel . . . . .   | 36        |
| 2.3.3    | Estimation du total dans un modèle bifactoriel sans interaction . . . . .                           | 39        |
| 2.3.4    | Estimation du total dans un modèle bifactoriel avec interaction . . . . .                           | 41        |
| 2.4      | Modèles avec variables continues . . . . .  | 44        |
| 2.4.1    | Modèle général avec covariance . . . . .  | 44        |
| 2.4.2    | Modèle unifactoriel avec une covariable . . . . .   | 46        |
| <b>3</b> | <b>Simulation</b> . . . . .   | <b>48</b> |
| 3.1      | Population . . . . .  | 48        |
| 3.1.1    | Première fonction : Loi Normale . . . . .   | 49        |
| 3.1.2    | Deuxième fonction : Processus Autorégressif d'ordre 1 . . . . .                                     | 49        |
| 3.1.3    | Troisième fonction : Processus Cyclique . . . . .   | 50        |
| 3.2      | Échantillonnage . . . . .   | 51        |
| 3.2.1    | Échantillonnage Aléatoire Simple . . . . .  | 51        |
| 3.2.2    | Échantillonnage Stratifié . . . . .   | 52        |
| 3.2.3    | Échantillonnage Non Standard . . . . .  | 52        |
| 3.3      | Inférence et résultats . . . . .  | 53        |
| 3.3.1    | Inférence . . . . .   | 53        |
| 3.3.2    | Résultats . . . . .   | 54        |
| 3.3.3    | Quand le modèle est "faux" . . . . .  | 56        |
|          | <b>Conclusion</b> . . . . .   | <b>59</b> |
|          | <b>A Données sur la population d'hôpitaux</b> . . . . .   | <b>61</b> |
|          | <b>B Construction d'un g-inverse par la méthode de décomposition en valeur singulière</b> . . . . . | <b>64</b> |
|          | <b>C Programmes R pour la simulation</b> . . . . .  | <b>66</b> |
| C.1      | La fonction population() . . . . .  | 66        |
| C.2      | La fonction tirage() . . . . .  | 68        |
| C.3      | La fonction inference() . . . . .   | 70        |
|          | <b>Bibliographie</b> . . . . .  | <b>80</b> |

# Liste des tableaux

|     |  |    |
|-----|--|----|
| 1.1 | <i>Données des N=33 hôpitaux de la population des 393 hôpitaux de l'annexe A . . . . .</i>                         | 6  |
| 1.2 | <i>Totaux dans la population des hôpitaux . . . . .</i>  | 8  |
| 2.1 | <i>Taille des cases formées par les niveaux des deux facteurs . . . . .</i>  | 41 |
| 3.1 | <i>forme en grille de la population à simuler . . . . .</i>  | 49 |
| 3.2 | <i>Vérification des paramètres de nos populations . . . . .</i>  | 55 |
| 3.3 | <i>Pourcentage des échantillons dans lesquels tous les effets sont estimés. . . . .</i>                            | 55 |
| 3.4 | <i>Moyennes des 10 Biais relatifs des estimateurs en pourcentage. . . . .</i>                                      | 56 |
| 3.5 | <i>Moyenne des 10 estimations de CV en pourcentage. . . . .</i>  | 56 |
| 3.6 | <i>Moyenne des 10 biais des estimateurs quand le modèle est "faux" en pourcentage. . . . .</i>                     | 57 |
| 3.7 | <i>p-value du test <math>H_0 : \text{biais} = 0</math> contre <math>H_1 : \text{biais} \neq 0</math> . . . . .</i> | 58 |
| 3.8 | <i>Moyenne des 10 estimations de CV en pourcentage quand le modèle est "faux". . . . .</i>                         | 58 |
| A.1 | <i>Données de la population de N=393 hôpitaux . . . . .</i>  | 63 |

# Table des figures

|     |  |    |
|-----|--|----|
| 1.1 | <i>Le nombre de sortis en fonction du nombre de lits . . . . .</i>                             | 7  |
| 1.2 | <i>Ajustement de la droite estimée . . . . .</i>   | 8  |
| 1.3 | <i>Échantillon "non équilibré" de la population par rapport à <math>X</math> . . . . .</i>     | 20 |
| 1.4 | <i>Échantillon "presque équilibré" de la population par rapport à <math>X</math> . . . . .</i> | 20 |

# Introduction

Depuis plusieurs siècles, l'Homme a recours à des types d'enquêtes pour se faire une idée sur certaines caractéristiques des populations. Les Égyptiens et les Romains pratiquaient déjà le recensement principalement pour obtenir des informations d'ordre militaire ou pour les taxes. Anders Kiaer, à la réunion de l'International Statistical Institute (ISI) en 1895, évoquait la possibilité d'utiliser un échantillon représentatif plutôt que toutes les données de la population pour les enquêtes sociales. Dans la première moitié du 20ème siècle, des scientifiques de divers domaines notamment en mathématiques, biologie, agriculture ... se sont intéressés à la théorie de l'échantillonnage. Dès lors des plans d'échantillonnage standards ont été développés. Devant l'absence d'outils informatiques, ces plans étaient très pratiques car la forme de la variance des estimateurs classiques (moyenne, total) était simple et bien connue. Cependant il existe de nombreuses situations où l'approche par le plan atteint rapidement ses limites. Nous pouvons penser au cas où le plan d'échantillonnage n'est pas standard avec un fort degré de complexité. Pour remédier à ces problèmes, une autre approche celle là basée sur les modèles de prédiction a fait son apparition.

Dans cette approche, la variable d'intérêt  $Y$  n'est pas considérée comme une quantité inconnue mais plutôt comme une variable aléatoire dont on veut prédire la valeur pour les individus non sélectionnés. Cette façon de voir la variable réponse est courante dans les autres domaines de la statistique. Il est dès lors possible de construire des modèles de prédictions et d'estimer les paramètres de ces modèles où des quantités comme la moyenne et le total. Pour ce faire les méthodes de régression, les modèles linéaires généraux ... bien familiers des statisticiens vont être utilisés.

Dans ce travail, il sera question de l'approche prédictive (par le modèle) dans la théorie de l'échantillonnage. Nous nous limiterons aux modèles linéaires généraux, ce qui représente une grande partie des modèles utilisés en pratique. Dans un premier temps, l'étude portera sur les modèles avec variables auxiliaires continues. Parmi les estimateurs non biaisés découlant du modèle, nous essayerons de trouver le prédicteur optimum c'est à dire celui à variance minimale. Dans la pratique, il est

souvent très difficile de connaître la loi exacte de la variable aléatoire d'intérêt  $Y$ . Il est de ce fait utile de savoir ce qui advient des estimateurs quand le modèle postulé ne représente pas bien les données.

Ensuite dans le chapitre 2, nous nous intéresserons aux modèles avec variables auxiliaires qualitatives (facteurs). La question est de savoir si la méthode fonctionne en cas de sur paramétrisation des niveaux des facteurs. La méthode de l'inverse généralisée y apportera une réponse positive.

Le dernier chapitre est consacré à une simulation dans un modèle avec deux variables qualitatives. Dans chaque cellule formée par les niveaux des facteurs, il y aura qu'une seule observation.

# Chapitre 1

## Modèles avec variables quantitatives

Dans la littérature, le livre "Finite Population Sampling and Inference A Prediction Approach" des auteurs Richard Valliant, Allan H. Dorfman et Richard M. Royall publié en 2000 est une bonne référence pour l'approche prédictive sous le modèle linéaire général. Ce livre consacre une bonne partie de ses chapitres à étudier les échantillons équilibrés sous l'approche prédictive. La revue de littérature sera pour l'essentiel basée sur ce livre.

### 1.1 Introduction à l'approche par le modèle

Dans Valliant, Dorfman et Royall (2000), la théorie est développée dans le contexte des populations finies. Une population finie est une collection d'unités distinctes qui peuvent être des personnes, des écoles, des ménages et bien d'autres choses. L'échantillonnage d'une population finie est une méthodologie qui consiste à choisir un sous ensemble de la population ou un ensemble d'unités de la population dans le but d'estimer des caractéristiques comme des moyennes ou des totaux pour la population entière. Dans les domaines de la statistique autres que l'échantillonnage, il est souvent question de variables aléatoires. Cependant l'approche classique par le plan d'échantillonnage ne traite pas d'une variable aléatoire dont on voudrait connaître la valeur future, mais bien d'une quantité existante que l'on essaye d'estimer, par exemple le revenu moyen, la production annuelle totale de pétrole. L'approche prédictive revient à la notion, la plus souvent rencontrée en statistique, celle de variable aléatoire.

### 1.1.1 Notation

La notation développée dans cette section sera utilisée pour l'ensemble du document. La population finie consiste en  $N$  unités distinctes et  $Y$  une caractéristique de la population qui nous intéresse, aussi appelée la variable d'intérêt. Le vecteur  $y = (y_1, \dots, y_N)$  est considéré comme la réalisation du vecteur aléatoire  $Y = (Y_1, \dots, Y_N)$ . Le but est d'estimer la quantité  $\theta = h(y_1, \dots, y_N)$ , une caractéristique de la population. Dans le cadre de ce travail la fonction  $h$  est une combinaison linéaire des  $y_i$  notée  $\gamma'y$ . Par exemple pour le total  $T = \sum_{i=1}^N y_i$  car  $\gamma' = (1, \dots, 1)'$  et pour la moyenne  $\mu = \sum_{i=1}^N y_i/N$  car  $\gamma' = (1/N, \dots, 1/N)'$ .

Dans la population, un échantillon  $s$  de taille  $n$  est tiré et les  $y_1, \dots, y_n$  sélectionnés sont observés. Pour la suite, l'indice  $s$  désignera l'échantillon et l'indice  $r$  désignera les  $N - n$  unités hors échantillon.  $\hat{\theta}_{opt}$  est le prédicteur linéaire non biaisé dont la variance est minimale, appelé le meilleur prédicteur linéaire non biaisé. Il faut considérer deux sortes de modèles : le modèle de travail noté  $M$  qui sert à toutes fins pratiques aux estimations et le modèle exact ou correct  $M^*$  qui est le meilleur modèle possible pour représenter la variable d'intérêt. Quand le modèle de travail  $M$  est différent du vrai modèle  $M^*$  alors le modèle  $M$  est dit faux.

Ce travail se limite aux modèles linéaires généraux. Pour des raisons pratiques les modèles considérés sont les modèles polynômiaux. En effet il est connu que la plupart des modèles continus peuvent être approximés par des polynômes par les techniques de linéarisation comme celle de Taylor. Le modèle polynômial général s'écrit

$$Y_i = \sum_{j=0}^J \delta_j \beta_j x_i^j + \varepsilon_i \gamma_i^{1/2}, \quad (1.1)$$

où les erreurs  $\varepsilon_i \sim (0, \sigma^2)$  et sont non corrélées, les  $\{\beta_j\}_{j=0}^J$  sont les paramètres inconnus à estimer et  $\{\delta_j\}_{j=0}^J = 0$  ou  $1$  et indique si le  $j$ ème terme est présent dans le modèle. La notation  $\varepsilon_i \sim (a, b)$  dénote une variable aléatoire avec une moyenne  $a$  et une variance  $b$ . Le modèle général (1.1) sera noté  $M(\delta_0, \dots, \delta_J : \gamma)$  et  $\hat{T}(\delta_0, \dots, \delta_J : \gamma)$  dénote le meilleur prédicteur linéaire non biaisé dérivé du modèle général (1.1) conformément à la notation de Royall et Hersen (1973a). Donc en particulier, la notation  $M(0, 1 : x)$  réfère au modèle  $Y_i = \beta_i x_i + \varepsilon_i x_i^{1/2}$  et  $\hat{T}(0, 1 : x)$  est l'estimateur du ratio qui coïncide avec le meilleur prédicteur linéaire non biaisé pour ce modèle trouvé dans la section (1.2.3). De même,  $M(1 : 1)$  réfère au modèle  $Y_i = \mu + \varepsilon_i$  et  $\hat{T}(1 : 1)$  est l'estimateur du total qui coïncide avec le meilleur prédicteur linéaire non biaisé pour le modèle trouvé dans la section 1.2.3.

### 1.1.2 Qu'est une approche par le modèle

L'approche classique par le plan d'échantillonnage considère une unité dans l'échantillon  $Y_s$  comme représentant un certain nombre d'unités  $Y_r$  non échantillonnées. Cela se manifeste par l'existence des probabilités de sélection ou encore des poids d'échantillonnage. Ainsi la fonction  $h_1(y_1, \dots, y_n)$  qui estime  $h(y_1, \dots, y_N)$  n'utilise que les  $Y_s$  observés dans l'échantillon en tenant compte du plan d'échantillonnage.

Pour ce qui est de l'approche par le modèle, les  $Y_i$  sont considérés comme des variables aléatoires. L'expression  $h(y_1, \dots, y_N)$  est une fonction conjointe des variables aléatoires  $Y_i$ . Donc l'estimation de  $h(y_1, \dots, y_N)$  revient à faire une prédiction pour les  $Y_r$  non échantillonnés c'est à dire non observés. Dans ce sens, le terme "prédiction" signifie "l'estimation statistique des  $Y_r$  non échantillonnés" et non "l'estimation de valeurs futures d'une variable aléatoire". En effet, les valeurs de  $Y$  existent déjà pour l'ensemble de la population. Cependant les valeurs de  $Y$  ne sont pas observées pour les unités hors échantillon.

Pour illustrer la construction des estimateurs prenons le total, dans ce cas

$$h(y_1, \dots, y_N) = T = \sum_{i=1}^N y_i.$$

Le total peut être divisé en deux, une partie observée et une deuxième non observée  $T = \sum_{i \in s} y_i + \sum_{i \in r} y_i$ . Pour estimer le total de la population, il suffit d'estimer le total des unités hors échantillon. Pour ce faire, chaque  $y_r$  est estimé comme réalisation de la variable aléatoire  $Y_r$  selon le modèle retenu. Considérons le modèle linéaire suivant :

$$Y_i = \beta x_i + \varepsilon_i \quad \text{où } E[\varepsilon_i] = 0 \text{ et } cov[\varepsilon_i, \varepsilon_j] = \begin{cases} x_i \sigma^2 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$$

Tous les  $y_r$  vont être estimés à partir du modèle linéaire et alors

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i = \sum_{i \in s} y_i + \sum_{i \in r} \hat{\beta} x_i$$

La théorie de la prédiction essaye donc de prédire mathématiquement les  $Y$  hors échantillon à partir d'un modèle dans le but d'avoir une estimation pour l'ensemble de la population.

### 1.1.3 Un exemple de l'approche par le modèle

Soit une population de  $N = 33$  hôpitaux. Soient  $Y$  le nombre de patients sortis et  $x$  le nombre de lits dans un hôpital.

| Unité | Numéro de l'hôpital | Nombre de lits (x) | Nombre de sorties (y) |
|-------|---------------------|--------------------|-----------------------|
| 1     | 44                  | 49                 | 173                   |
| 2     | 45                  | 50                 | 260                   |
| 3     | 62                  | 64                 | 225                   |
| 4     | 71                  | 70                 | 209                   |
| 5     | 91                  | 99                 | 346                   |
| 6     | 94                  | 100                | 383                   |
| 7     | 95                  | 100                | 318                   |
| 8     | 96                  | 100                | 373                   |
| 9     | 127                 | 128                | 577                   |
| 10    | 168                 | 184                | 381                   |
| 11    | 187                 | 224                | 590                   |
| 12    | 192                 | 227                | 732                   |
| 13    | 196                 | 231                | 931                   |
| 14    | 197                 | 233                | 683                   |
| 15    | 203                 | 244                | 858                   |
| 16    | 214                 | 260                | 1076                  |
| 17    | 226                 | 275                | 1201                  |
| 18    | 230                 | 279                | 754                   |
| 19    | 245                 | 303                | 715                   |
| 20    | 251                 | 309                | 985                   |
| 21    | 269                 | 347                | 1166                  |
| 22    | 272                 | 350                | 1173                  |
| 23    | 286                 | 373                | 787                   |
| 24    | 303                 | 411                | 808                   |
| 25    | 304                 | 417                | 1369                  |
| 26    | 314                 | 451                | 1584                  |
| 27    | 337                 | 523                | 1232                  |
| 28    | 348                 | 549                | 1547                  |
| 29    | 351                 | 551                | 1645                  |
| 30    | 353                 | 558                | 1152                  |
| 31    | 354                 | 562                | 2116                  |
| 32    | 361                 | 591                | 999                   |
| 33    | 393                 | 986                | 2268                  |

TAB. 1.1 – Données des  $N=33$  hôpitaux de la population des 393 hôpitaux de l'annexe A

Le tableau (1.1) illustre les données pour toute la population. Le but est d'estimer le nombre total  $T = \sum_{i=1}^{33} y_i$  de patients sortis des 33 hôpitaux. L'information sur la variable auxiliaire  $x$  est connue pour toute la population.

Considérons un échantillon de  $n = 32$  hôpitaux, seulement l'hôpital numéro 16 n'a

pas été tiré. Le total sera donc la somme du nombre de patients des 32 hôpitaux échantillonnés ( $y_s$ ) plus le nombre  $y_{16}$  de patients sortis pour l'hôpital numéro 16 qui ne se trouve pas dans l'échantillon. Ainsi il vient que  $T = \sum_{i \in s} y_i + y_{16}$ . A cette étape, il faut trouver un modèle qui estime  $y_{16}$ . Pour cela, traçons le graphe des  $y_i$  en fonction des  $x_i$  pour les 32 hôpitaux.

**Graphique du nombre de sortis en fonction du nombre de lits**

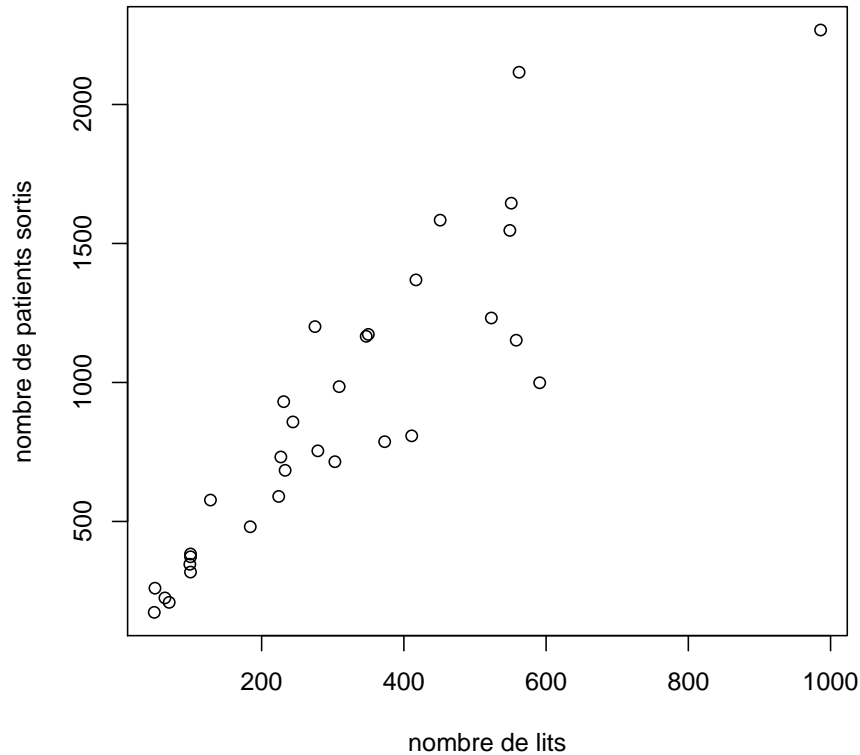


FIG. 1.1 – Le nombre de sortis en fonction du nombre de lits

La figure 1.1 montre qu'une relation linéaire entre  $Y$  et  $X$  est très raisonnable. L'ordonnée à l'origine peut être considérée nulle. Le modèle retenu sera donc :

$$Y_i = \beta x_i + \varepsilon_i,$$

où  $E[\varepsilon_i] = 0$  et  $cov[\varepsilon_i, \varepsilon_j] = x_i \sigma^2$  si  $i = j$  et 0 sinon. Le coefficient de la régression  $\beta$  sera donc estimé par  $\hat{\beta}$  qui minimise la somme des carrés pondérés

$$g(\beta) = \sum_{i \in s} \frac{1}{\sigma^2 x_i} (y_i - \beta x_i)^2.$$

La méthode classique de minimisation donne  $\hat{\beta} = \sum_{i \in s} y_i / \sum_{i \in s} x_i$ . En remplaçant par les données du tableau 1.1, nous obtenons  $\hat{\beta} = 28641/9938 \simeq 2.88$ . La prévision du

nombre de patients sortis est  $\hat{y}_{16} = \hat{\beta}x_{16} \simeq 2.88 \times 260$  soit  $\hat{y}_{16} \simeq 748.8 \simeq 749$  patients. La figure 1.2 montre l'ajustement de droite estimée dans le nuage de points.

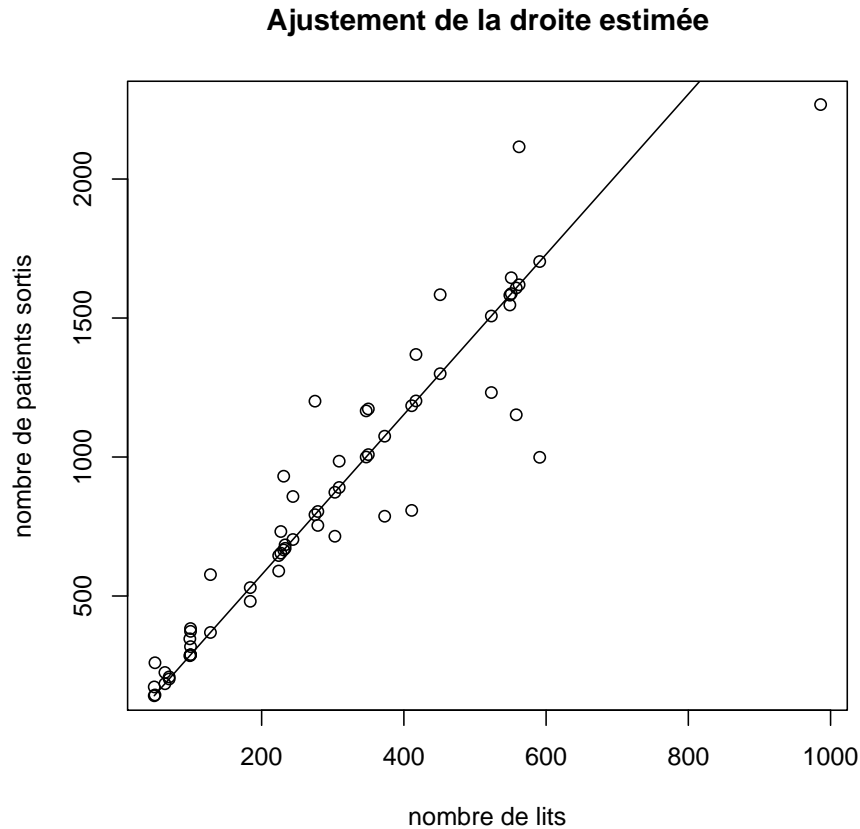


FIG. 1.2 – *Ajustement de la droite estimée*

Le tableau (1.1) donne les totaux suivants :

| Unité               | Nombre de lits (x) | Nombre de sorties (y) |
|---------------------|--------------------|-----------------------|
| Total pour les N=33 | 10198              | 29717                 |
| Total pour les n=32 | 9938               | 28641                 |

TAB. 1.2 – *Totaux dans la population des hôpitaux*

D'où l'estimation pour le total :  $\hat{T} = \sum_{i \in s} y_i + \hat{y}_{16} = 28641 + 749 = 29390$

Par rapport à la vraie valeur du total, la méthode commet une erreur relative de  $|\hat{T} - T|/T = |29390 - 29717|/29717 = 0.011$ . Donc l'erreur relative est de l'ordre de 1%.

### 1.1.4 Pourquoi une approche par le modèle

Tout d'abord, il n'y a aucun doute que les deux théories sont mathématiquement valides. L'approche classique par le plan d'échantillonnage est une méthode très couramment utilisée. Quand le plan est très complexe, il y a un recours à des estimateurs qui peuvent se révéler assez imprécis. De plus, très souvent, il faut tenir compte de contraintes d'ordre budgétaire, environnemental, conditions de travail et autres sur le plan d'enquête. Dans certains cas, ces contraintes font qu'il est impossible d'appliquer un plan d'échantillonnage classique comme celui aléatoire simple ou encore par grappes etc. Or la méthodologie de calculs des estimateurs est basée sur les plans d'échantillonnage. Si l'enquête ne suit pas un plan d'échantillonnage connu, les estimateurs existant ne seront pas optimaux, voir même adaptés.

Dans de telles situations, l'approche par le modèle est recommandée. Puisqu'il n'est pas absolument indispensable d'avoir un plan d'échantillonnage connu. L'approche prédictive considère la variable d'intérêt comme une variable aléatoire. Il s'agit alors de faire des prévisions pour les unités hors échantillon par rapport au modèle. C'est donc une approche probabiliste classique. La théorie des probabilités étant maîtrisée et bien connue, l'approche par le modèle s'impose naturellement. Le fait qui peut paraître comme un inconvénient majeur est l'éventuelle inexactitude du modèle. Le modèle de travail choisi peut ne pas être le vrai modèle suivi par la variable d'intérêt. Dans ce cas, les estimations restent-elles valables ? Il existe des façons de se protéger des modèles inexacts notamment avec les échantillons équilibrés. Mais avant de voir ce cas, la prochaine partie va être consacrée à étudier l'approche par le modèle dans le cas des modèles linéaires.

## 1.2 L'approche prédictive sous le modèle linéaire général

L'approche prédictive, introduite précédemment, est appliquée ici dans le cas des modèles linéaires généraux. Le théorème général de la prédiction (Royall 1976b) permet de déduire de ces modèles le meilleur estimateur non biaisé et sa variance. Cela pour n'importe quelle caractéristique qui soit une combinaison linéaire de la variable d'intérêt comme le total ou la moyenne. Le théorème général de la prédiction sera appliqué pour quelques modèles linéaires simples afin d'estimer des totaux, des proportions. Ensuite pour finir, des intervalles de confiance seront construits pour les meilleurs prédicteurs non biaisés.

### 1.2.1 Définitions

Pour tout échantillon tiré, il est possible de réorganiser le vecteur des observations comme suit  $y = (y_s, y_r)$  où  $y_s$  est le sous vecteur des  $n$  unités de l'échantillon et  $y_r$  celui des  $N - n$  unités hors échantillon. La quantité à estimer peut alors s'écrire  $\gamma'y = \gamma'(y_s, y_r) = \gamma'_s y_s + \gamma'_r y_r$ . Cette dernière égalité montre qu'estimer  $\gamma'y$  revient à prédire  $\gamma'_r y_r$ . Les estimateurs utilisés dans ce chapitre sont définis comme suit :

**Définition 1.** Un estimateur linéaire de  $\theta = \gamma'Y$  est défini par  $\hat{\theta} = g'_s Y_s$  où  $g'_s = (g_1, \dots, g_n)'$  est un vecteur coefficient de longueur  $n$ .

**Définition 2.** L'erreur de l'estimateur  $\hat{\theta} = g'_s Y_s$  est définie par  $\hat{\theta} - \theta = g'_s Y_s - \gamma'Y$ .

L'erreur de l'estimateur peut être réécrite en terme des unités dans l'échantillon et hors de l'échantillon.

$$\begin{aligned} g'_s Y_s - \gamma'Y &= (g'_s - \gamma'_s)Y_s - \gamma'_r Y_r \\ &= a'Y_s + \gamma'_r Y_r, \end{aligned}$$

où  $a = g_s - \gamma_s$ . Ici le premier terme dépend de l'échantillon. Le second terme dépend quant à lui des unités hors échantillon et donc doit être prédit. Donc utilisé  $g'_s Y_s$  pour estimer  $\gamma'Y$  est équivalent à prendre  $a'Y_s$  pour prédire  $\gamma'_r Y_r$ . En d'autres termes, trouver le meilleur  $g_s$  pour l'estimation est équivalent à trouver le meilleur  $a$ .

Le problème de la prédiction est étudié sous le modèle linéaire général  $M$  défini comme suit :

$$\begin{aligned} E_M(Y) &= X\beta \\ \text{var}_M(Y) &= V, \end{aligned} \tag{1.2}$$

où  $X$  est une matrice  $N \times p$ ,  $\beta$  est un vecteur  $p \times 1$  de paramètres inconnus, et  $V$  est une matrice de variance-covariance définie positive. La variable auxiliaire  $X$  est supposée connue pour chaque unité de la population. Il est possible de réarranger les unités de la population de telle façon que les unités dans l'échantillon soient les  $n$  premiers éléments de  $Y$ , ce qui donne :

$$X = \begin{bmatrix} X_s \\ X_r \end{bmatrix}, \quad V = \begin{bmatrix} V_{ss} & V_{sr} \\ V_{rs} & V_{rr} \end{bmatrix},$$

où  $X_s$  est de dimension  $n \times p$ ,  $X_r$  est  $(N-n) \times p$ ,  $V_{ss}$  est  $n \times n$ ,  $V_{rr}$  est  $(N-n) \times (N-n)$ ,  $V_{sr}$  est  $n \times (N-n)$  et  $V_{rs} = V'_{sr}$ .  $V_{ss}$  est par hypothèse définie positive.

Définissons la notion de "non biaisé" et la "variance de l'erreur".

**Définition 3.** L'estimateur  $\hat{\theta}$  est non biaisé pour  $\theta$  sous le modèle linéaire  $M$  si  $E_M(\hat{\theta} - \theta) = 0$ .

Les expressions "prédiction non biaisée" ou "modèle non biaisé" peuvent être utilisées pour parler d'un estimateur non biaisé sous le modèle.

**Définition 4.** La variance de l'erreur de l'estimateur non biaisé  $\hat{\theta}$  sous le modèle  $M$  est  $E_M(\hat{\theta} - \theta)^2$ .

### 1.2.2 Théorème général de la prédiction (Royall 1976b)

Ce théorème permet, sous un modèle linéaire, d'obtenir parmi la classe de prédicteurs non biaisés celui qui a la variance minimale.

**Théorème 1** (Royall 1976b). Parmi les prédicteurs linéaires non biaisés  $\hat{\theta}$  de  $\theta$ , la variance de l'erreur est minimisée par

$$\hat{\theta}_{opt} = \gamma'_s Y_s + \gamma'_r [X_r \hat{\beta} + V_{rs} V_{ss}^{-1} (Y_s - X_s \hat{\beta})], \quad (1.3)$$

où

$$\hat{\beta} = (X'_s V_{ss}^{-1} X_s)^{-1} X'_s V_{ss}^{-1} Y_s.$$

La variance de l'erreur de  $\hat{\theta}_{opt}$  est :

$$\begin{aligned} var_M(\hat{\theta}_{opt} - \theta) &= \gamma'_r (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r \\ &+ \gamma'_r (X_r - V_{rs} V_{ss}^{-1} X_s) (X'_s V_{ss}^{-1} X_s)^{-1} (X_r - V_{rs} V_{ss}^{-1} X_s)' \gamma_r. \end{aligned} \quad (1.4)$$

#### Preuve du théorème général

Notons d'abord que

$$\begin{aligned} E_M(g'_s Y_s - \gamma'_r Y_r)^2 &= E_M(a' Y_s - \gamma'_r Y_r)^2 \\ &= var_M(a' Y_s - \gamma'_r Y_r) + [E_M(a' Y_s - \gamma'_r Y_r)]^2 \\ &= a' V_{ss} a - 2a' V_{sr} \gamma_r + \gamma'_r V_{rr} \gamma_r + [(a' X_s - \gamma'_r X_r) \beta]^2 \end{aligned} \quad (1.5)$$

où  $a' X_s \beta$  estime  $\gamma'_r X_r \beta$  sans biais donc  $[(a' X_s - \gamma'_r X_r) \beta]^2 = 0$ .

Le but de la preuve est de minimiser la fonction (1.5) par rapport à  $a$ . Le minimum est la variance optimale (1.4) qui correspond à l'estimateur (1.3). Pour trouver le minimum en tenant compte de la contrainte du biais nul, la méthode du lagrangien sera utilisée

$$L_\lambda(a) = a' V_{ss} a - 2a' V_{sr} \gamma_r + 2(a' X_s - \gamma'_r X_r) \lambda \quad (1.6)$$

La fonction (1.6) est le lagrangien et  $\lambda^* = 2\lambda$  est le multiplicateur de Lagrange, le nombre 2 est mis en évidence pour la commodité du calcul. Le terme  $\gamma_r' V_{rr} \gamma_r$  du membre de droite n'est pas considéré dans le lagrangien car une constante n'influence pas le calcul du minimum.

Le vecteur des dérivées partielles de  $L_\lambda(a)$  par rapport à  $a = g_s - \gamma_s$  est :

$$\partial L_\lambda(a)/\partial a = 2V_{ss}a - 2V_{sr}\gamma_r + 2X_s\lambda$$

L'équation  $\partial L_\lambda(a)/\partial a = 0$  s'écrit :

$$X_s\lambda = V_{sr}\gamma_r - V_{ss}a \quad (1.7)$$

$$a = V_{ss}^{-1}(V_{sr}\gamma_r - X_s\lambda) \quad (1.8)$$

Multiplions à gauche l'équation (1.7) par le vecteur  $X_s'V_{ss}^{-1}$ , nous obtenons :

$$X_s'V_{ss}^{-1}X_s\lambda = X_s'V_{ss}^{-1}V_{sr}\gamma_r - X_s'V_{ss}^{-1}V_{ss}a$$

or  $a'X_s = \gamma_r'X_r$  ce qui donne  $X_s'a = X_r'\gamma_r$ , d'où

$$\begin{aligned} \lambda &= [X_s'V_{ss}^{-1}X_s]^{-1}[X_s'V_{ss}^{-1}V_{sr}\gamma_r - X_r'\gamma_r] \\ \lambda &= A^{-1}(X_s'V_{ss}^{-1}V_{sr} - X_r')\gamma_r \end{aligned} \quad (1.9)$$

où  $A = [X_s'V_{ss}^{-1}X_s]$

On remplace  $\lambda$  par son expression (1.9) dans l'équation (1.8)

$$\begin{aligned} a_{opt} &= V_{ss}^{-1}[V_{sr}\gamma_r - X_sA^{-1}(X_s'V_{ss}^{-1}V_{sr} - X_r')\gamma_r] \\ a_{opt} &= V_{ss}^{-1}[V_{sr} - X_sA^{-1}(X_s'V_{ss}^{-1}V_{sr} - X_r')]\gamma_r \\ a_{opt} &= V_{ss}^{-1}[V_{sr} + X_sA^{-1}(X_r' - X_s'V_{ss}^{-1}V_{sr})]\gamma_r \end{aligned} \quad (1.10)$$

Il ne reste plus qu'à remplacer  $a$  par  $a_{opt}$  dans l'expression  $\hat{\theta} = \gamma_s'Y_s + a'Y_s$ .

$$\begin{aligned} \hat{\theta}_{opt} &= \gamma_s'Y_s + \gamma_r'[V_{rs} + (X_r - V_{rs}V_{ss}^{-1}X_s)A^{-1}X_s']V_{ss}^{-1}Y_s \\ \hat{\theta}_{opt} &= \gamma_s'Y_s + \gamma_r'[V_{rs}V_{ss}^{-1}Y_s + X_rA^{-1}X_s'V_{ss}^{-1}Y_s - V_{rs}V_{ss}^{-1}X_sA^{-1}X_s'V_{ss}^{-1}Y_s] \\ \hat{\theta}_{opt} &= \gamma_s'Y_s + \gamma_r'[X_r\hat{\beta} + V_{rs}V_{ss}^{-1}(Y_s - X_s\hat{\beta})] \end{aligned}$$

On retrouve ainsi l'expression (1.3) du meilleur prédicteur linéaire non biaisé du théorème général. De même, la variance minimale est :

$$var_M(\hat{\theta}_{opt} - \theta) = a_{opt}'V_{ss}a_{opt} - 2a_{opt}'V_{sr}\gamma_r + \gamma_r'V_{rr}\gamma_r$$

Notons que :

$$a_{opt}'V_{ss} = \gamma_r'[V_{rs} + (X_r - V_{rs}V_{ss}^{-1}X_s)A^{-1}X_s']$$

Ainsi :

$$\begin{aligned}
a'_{opt} V_{ss} a_{opt} &= \gamma'_r [V_{rs} + (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s] V_{ss}^{-1} [V_{sr} + X_s A^{-1} (X'_r - X'_s V_{ss}^{-1} V_{sr})] \gamma_r \\
&= \gamma'_r [V_{rs} V_{ss}^{-1} + (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s V_{ss}^{-1}] [V_{sr} + X_s A^{-1} (X'_r - X'_s V_{ss}^{-1} V_{sr})] \gamma_r \\
&= \gamma'_r [V_{rs} V_{ss}^{-1} V_{sr} + V_{rs} V_{ss}^{-1} X_s A^{-1} (X'_r - X'_s V_{ss}^{-1} V_{sr}) + (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} \times \\
&\quad X'_s V_{ss}^{-1} V_{sr} + (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s V_{ss}^{-1} X_s A^{-1} (X'_r - X'_s V_{ss}^{-1} V_{sr})] \gamma_r
\end{aligned}$$

À droite de l'égalité, nous avons une fonction quadratique d'où :  $(X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s V_{ss}^{-1} V_{sr} + X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s V_{ss}^{-1} V_{sr} = 2(X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s V_{ss}^{-1} V_{sr}$ .

De plus  $A^{-1} X'_s V_{ss}^{-1} X_s = A^{-1} \times A = I$  où  $I$  est la matrice identité. Nous obtenons donc :

$$\begin{aligned}
a'_{opt} V_{ss} a_{opt} &= \gamma'_r [V_{rs} V_{ss}^{-1} V_{sr} + 2(X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s V_{ss}^{-1} V_{sr} \\
&\quad + (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} (X'_r - X'_s V_{ss}^{-1} V_{sr})] \gamma_r
\end{aligned}$$

$$\begin{aligned}
a'_{opt} V_{sr} &= \gamma'_r [V_{rs} + (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s] V_{ss}^{-1} V_{sr} \\
&= \gamma'_r [V_{rs} V_{ss}^{-1} V_{sr} + (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} X'_s V_{ss}^{-1} V_{sr}]
\end{aligned}$$

Finalement, après simplification et regroupement, nous obtenons :

$$\begin{aligned}
E_M(\hat{\theta}_{opt} - \theta)^2 &= var_M(\hat{\theta}_{opt}) \\
&= \gamma'_r (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r \\
&\quad + \gamma'_r (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} (X'_r - X'_s V_{ss}^{-1} V_{sr}) \gamma_r \\
&= \gamma'_r (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r \\
&\quad + \gamma'_r (X_r - V_{rs} V_{ss}^{-1} X_s) A^{-1} (X_r - V_{rs} V_{ss}^{-1} X_s)' \gamma_r
\end{aligned}$$

✠

Remarquons deux points :

- L'estimateur  $\hat{\beta}$  est la solution de la minimisation de la somme des carrés pondérés  $(Y_s - X_s \beta)' V_{ss}^{-1} (Y_s - X_s \beta)$  par rapport à  $\beta$ . La solution donne :

$$\begin{aligned}
X'_s V_{ss}^{-1} X_s \hat{\beta} &= X'_s V_{ss}^{-1} Y_s \\
\hat{\beta} &= (X'_s V_{ss}^{-1} X_s)^{-1} X'_s V_{ss}^{-1} Y_s
\end{aligned}$$

quand  $X'_s V_{ss}^{-1} X_s$  est inversible.

- Quand  $X$  est une variable qualitative prenant les valeurs 0 et 1, il arrive que la matrice  $X'_s V_{ss}^{-1} X_s$  ne soit pas inversible. Dans ce cas le théorème général ne s'applique pas sous cette forme. Mais il est possible de considérer des modèles de régression tels que  $X'_s V_{ss}^{-1} X_s$  soit singulière. Dans ce cas  $\hat{\beta}$  admet plusieurs solutions et des estimateurs qui tiennent compte de cette non unicité existent ( chapitre 7 de Valliant, Dorfman et Royall (2000)).

Quand les  $Y_i$  ne sont pas corrélés, les résultats sont beaucoup plus simples. D'où le corollaire

**Corollaire 1.** *Sous le modèle (1.2), si les  $Y_i$  ne sont pas corrélés c'est à dire que  $V_{rs} = V_{sr} = 0$  alors le meilleur prédicteur linéaire non biaisé est  $\hat{\theta}_{opt} = \gamma'_s Y_s + \gamma'_r X_r \hat{\beta}$  et la variance de l'erreur est  $var_M(\hat{\theta}_{opt} - \theta) = \gamma'_r (V_{rr} + X_r A^{-1} X'_r) \gamma_r$*

### 1.2.3 Le meilleur prédicteur linéaire non-biaisé sous quelques modèles simples

La caractéristique à estimer est le total  $\theta_{opt} = \sum_{i=1}^N Y_i = T$  donc  $\gamma = (1, \dots, 1)'$ . Dans les exemples ci-dessous, le théorème (1) permet de déterminer le meilleur prédicteur linéaire non biaisé. Dans les modèles considérés, les  $Y_i$  seront non corrélés donc :  $\hat{\theta}_{opt} = \sum_s Y_s + \sum_r X_r \hat{\beta}$ . Dans certains cas simples, le meilleur prédicteur linéaire non biaisé coïncide avec l'estimateur classique de la méthode basé sur le plan d'échantillonnage.

#### Exemple 1 : Estimateur du total par la moyenne

Le modèle est :

$$y_i = \mu + \varepsilon_i \quad \text{où} \quad \varepsilon_i \sim (0, \sigma^2) \quad \text{non corrélés}$$

Nous avons donc :  $X = (1, \dots, 1)'$ ,  $\beta = \mu$  et  $V = \sigma^2 I_N$

Le meilleur prédicteur est :  $\hat{\theta}_{opt} = \sum_s Y_i + \sum_r \hat{\beta}$  où

$$\begin{aligned} \hat{\beta} &= (X'_s V_{ss}^{-1} X_s)^{-1} X'_s V_{ss}^{-1} Y_s \\ &= \sigma^2 (X'_s I_N X_s)^{-1} \frac{1}{\sigma^2} X'_s Y_s \\ &= \frac{1}{n} X'_s Y_s \quad \text{car} \quad X'_s X_s = (1, \dots, 1) * (1, \dots, 1)' = n \\ &= \frac{1}{n} \sum_{i \in s} Y_s = \bar{Y}_s \end{aligned}$$

Donc il vient que :  $\hat{\theta}_{opt} = \sum_s Y_i + \sum_r \bar{Y}_s = n \bar{Y}_s + r \bar{Y}_s = N \bar{Y}_s$ .

D'après le corollaire 1,  $var_M(\hat{\theta}_{opt} - \theta) = \gamma'_r (V_{rr} + X_r A^{-1} X'_r) \gamma_r$ .

D'où

$$\begin{aligned}
 \text{var}_M(\hat{\theta}_{opt} - \theta) &= \gamma_r'(V_{rr} + X_r A^{-1} X_r') \gamma_r \\
 &= \gamma_r' \left( \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} + \begin{pmatrix} \sigma^2/n & \cdots & \sigma^2/n \\ \vdots & \ddots & \vdots \\ \sigma^2/n & \cdots & \sigma^2/n \end{pmatrix} \right) \gamma_r \\
 &= \sigma^2 \gamma_r' \begin{pmatrix} 1 + 1/n & 1/n & 1/n \\ 1/n & \ddots & 1/n \\ 1/n & 1/n & 1 + 1/n \end{pmatrix} \gamma_r
 \end{aligned}$$

$$\begin{aligned}
 \text{var}_M(\hat{\theta}_{opt} - \theta) &= \sigma^2(r + r^2/n) \\
 &= \sigma^2 r(1 + r/n) = \sigma^2(N - n)(N/n) \quad \text{car } r = N - n \\
 &= N^2(N - n)/N\sigma^2/n
 \end{aligned}$$

La variance de l'erreur du meilleur estimateur linéaire non biaisé pour ce modèle est :

$$\text{var}_M(\hat{\theta}_{opt} - \theta) = N^2(1 - f)\sigma^2/n, \quad \text{où } f = n/N$$

### Exemple 2 : Estimateur du total par la régression linéaire

Le modèle est :

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{où } \varepsilon_i \sim (0, \sigma^2) \quad \text{non corrélés}$$

Le meilleur prédicteur est :  $\hat{\theta}_{opt} = \sum_s Y_i + \sum_r \hat{\beta} x_i$  où  $\hat{\beta} = (X_s' V_{ss}^{-1} X_s)^{-1} X_s' V_{ss}^{-1} Y_s$

Les méthodes habituelles permettent de dériver les estimateurs suivants :

$$\hat{\beta}_1 = \frac{\sum_s (y_i - \bar{y}_s)(x_i - \bar{x}_s)}{\sum_s x_i^2 - n\bar{x}_s^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{y}_s - \hat{\beta}_1 \bar{x}_s.$$

Donc il vient que :

$$\begin{aligned}
\hat{\theta}_{opt} &= \sum_s Y_i + \sum_r (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= n\bar{Y}_s + \sum_r (\bar{y}_s - \hat{\beta}_1 \bar{x}_s + \hat{\beta}_1 x_i) \\
&= n\bar{Y}_s + \sum_r \bar{y}_s + \hat{\beta}_1 \sum_r (x_i - \bar{x}_s) \\
&= N\bar{Y}_s + \hat{\beta}_1 (\sum_r x_i - \sum_r \bar{x}_s) \\
&= N\bar{Y}_s + \hat{\beta}_1 (\sum_r x_i + \sum_s x_i - \sum_r \bar{x}_s - \sum_s x_i) \\
&= N\bar{Y}_s + \hat{\beta}_1 (\sum_N x_i - r\bar{x}_s - s\bar{x}_s) \\
&= N\bar{Y}_s + \hat{\beta}_1 N(\bar{x} - \bar{x}_s) \\
\hat{\theta}_{opt} &= N(\bar{Y}_s + \hat{\beta}_1(\bar{x} - \bar{x}_s))
\end{aligned}$$

La variance de l'erreur de l'estimateur pour la regression linéaire est :

$$var_M(\hat{\theta}_{opt} - \theta) = N^2(1-f)\sigma^2[1 + (\bar{x}_s - \bar{x})^2 / \{(1-f)c_s\}] / n, \quad \text{où } c_s = \sum_s (x_i - \bar{x}_s)^2 / n.$$

### Exemple 3 : Estimateur stratifié du total

Cet exemple est très semblable à l'exemple 1. En fait dans une strate donnée le modèle est celui de l'exemple 1. D'où le modèle global est :

$$y_{hi} = \mu_h + \varepsilon_{hi} \quad \text{où} \quad \varepsilon_{hi} \sim (0, \sigma^2) \quad \text{non corrélés}$$

La même démonstration qu'à l'exemple 1 montre que :

- le total dans une strate h s'écrit  $\hat{\theta}_{hopt} = N_h \bar{Y}_{hs}$  où  $N_h$  est la taille de la strate h et  $\bar{Y}_{hs}$  est la moyenne dans la strate h. Donc le total global est la somme des totaux dans les H strates  $\hat{\theta}_{opt} = \sum_h N_h \bar{Y}_{hs}$ .
  - la variance dans une strate h s'écrit  $var_M(\hat{\theta}_{hopt} - \theta_h) = N_h^2(1-f_h)\sigma_h^2/n_h$ , où  $f_h = n_h/N_h$  et  $n_h$  est la taille de l'échantillon dans la strate h.
- La variance de l'erreur du meilleur estimateur linéaire non biaisé pour ce modèle est :  $var_M(\hat{\theta}_{opt} - \theta) = \sum_h N_h^2(1-f_h)\sigma_h^2/n_h$ .

**Exemple 4 : Estimateur du total par le modèle du ratio**

Le modèle est :

$$y_i = \beta x_i + \varepsilon_i x_i \quad \text{où} \quad \varepsilon_i \sim (0, \sigma^2) \quad \text{non corrélés}$$

Le meilleur prédicteur est :  $\hat{\theta}_{opt} = \sum_s Y_i + \sum_r \hat{\beta} x_i$  où

$$\hat{\beta} = (X'_s V_{ss}^{-1} X_s)^{-1} X'_s V_{ss}^{-1} Y_s$$

$$\begin{aligned} \hat{\beta} &= [(x_1 \dots x_n) \begin{pmatrix} \sigma^2 x_1^2 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \sigma^2 x_n^2 \end{pmatrix}^{-1} (x_1 \dots x_n)']^{-1} X'_s V_{ss}^{-1} Y_s \\ &= \sigma^2 (1/x_1 \dots 1/x_n) (x_1 \dots x_n)' X'_s V_{ss}^{-1} Y_s \\ &= \sigma^2 / n X'_s V_{ss}^{-1} Y_s \\ &= \sigma^2 / n (1/\sigma^2 x_1 \dots 1/\sigma^2 x_n) Y_s \\ &= 1/n (1/x_1 \dots 1/x_n) (y_1 \dots y_n)' \\ &= \sum_s y_i / (n x_i) \end{aligned}$$

La variance de l'erreur de l'estimateur par le modèle du ratio est :

$$var_M(\hat{\theta}_{opt} - \theta) = \sigma^2 [(N - n)^2 \bar{x}_r^2 / n + \sum_r x_i^2]$$

**Exemple 5 : Estimation d'une proportion**

Le modèle est :

$$Y_i = \begin{cases} 1 & \text{avec} \quad P(Y_i = 1) = p \\ 0 & P(Y_i = 0) = q = 1 - p \end{cases}$$

Sous ce modèle, nous avons que :  $E_M(Y_i) = p$  et  $var_M(Y_i) = pq$ .

On veut prédire  $p = \sum_{i=1}^N Y_i / N$ .

Le meilleur prédicteur est :

$$\begin{aligned} \hat{\theta}_{opt} &= \gamma'_s Y_s + 0 \\ &= (1/n \dots 1/n) (y_1 \dots y_n)' \\ &= \sum_s y_i / n \end{aligned}$$

La variance de l'erreur de l'estimateur de la proportion est :

$$\text{var}_M(\hat{\theta}_{opt} - \theta) = (1 - f)pq/n.$$

## 1.3 Approche prédictive et échantillons équilibrés

Jusqu'ici un modèle linéaire a été considéré pour la variable d'intérêt. Les prédictions et les estimations ont été dérivées à partir de ce modèle. Il est légitime alors de s'interroger sur les propriétés des estimateurs si le modèle postulé est faux. En d'autres termes si le modèle considéré ne représente pas bien les données. Le cas échéant les estimations de la caractéristique, par exemple le total, restent elles valides ? Cette section montre que la sélection d'un certain type d'échantillons appelés "échantillons équilibrés" et de prédicteurs appropriés permettent de rendre nos estimateurs robustes. Robuste voulant dire ici que le biais de l'estimateur est robuste à la considération d'un modèle faux. Comme dans ce qui précède, la caractéristique principale, sur laquelle l'emphase sera mise, est le total. Deux modèles sont à l'étude le vrai modèle  $M^*$  et le modèle de travail  $M$ . Le modèle de travail est dit faux s'il est différent du vrai modèle qui représente le mieux les données.

### 1.3.1 Robustesse et biais

En général dans la littérature, un estimateur est dit robuste s'il n'est pas beaucoup affecté par des changements dans le mécanisme (modèle) par lequel les données sont générées (voir Huber 1981, section 1.1-2).

Dans l'approche classique par le plan d'échantillonnage, la robustesse est défini autrement. En effet les données sont considérées comme fixes, car elles n'ont pas besoin d'un modèle pour être générées. Dans ce contexte, les estimations sont toujours robustes n'étant pas altérées par un modèle faux. A ce propos Brewer et Sarndal (1983) écriront une phrase dont une traduction en français pourrait être :

*les méthodes d'échantillonnage probabilistes sont robustes par définition, puisqu'elles ne font pas appel à un modèle, il n'y a pas besoin de discuter de l'impact d'utiliser un modèle faux.*

La robustesse sera donc généralement considérée comme étant la capacité des estimations d'être insensibles c'est à dire de rester valides quand les hypothèses sont violées. En pratique pour s'assurer de cette robustesse :

- on choisit notre échantillon de façon aléatoire, c'est la randomisation.

- on utilise des estimateurs qui tiennent compte de la nature de nos données.
- on utilise une grande taille d'échantillon afin de se servir du théorème central limite.

A ce propos Hansen, Madow et Tepping (p. 791, 1983) écrivaient :

*c'est avantageux ... et suffisant d'avoir un "bon" estimateur basé sur une taille raisonnablement grande de l'échantillon aléatoire qui produit un intervalle de confiance valide*

Dans l'approche par la prédiction, la notion de robustesse retrouve son sens premier. En effet, sous cette approche, le modèle de travail  $M$  est souvent une simplification du vrai modèle  $M^*$  qui a généré les données ou qui représente le mieux les données. La question qui se pose est de savoir comment se comporte le biais de l'estimateur lorsque le modèle de travail  $M$  n'est pas le meilleur modèle. D'aucun pourrait se dire pourquoi ne pas simplement utiliser le meilleur prédicteur linéaire non biaisé dérivé du vrai modèle  $M^*$ ? Tout simplement parce qu'en pratique le vrai modèle  $M^*$  n'est souvent pas connu. Il est toujours possible de chercher d'autres modèles qui représentent mieux les données, tout en faisant un compromis entre la complexité du modèle et la qualité des prédicteurs dérivés de ce modèle.

En pratique, après avoir obtenu les données, il faut trouver un modèle qui représente bien la variable d'intérêt en fonction de l'information auxiliaire disponible  $X$ . C'est ce modèle trouvé qui est appelé le modèle de travail. Le meilleur prédicteur linéaire non biaisé  $\hat{\theta}_{opt}$  est calculé à partir de ce modèle pour estimer  $\theta$ . Le but est de s'assurer que même si notre modèle de travail est faux, c'est à dire qu'il existe un meilleur modèle pour représenter nos données, nos estimations restent non biaisées ou presque non biaisées. Illustrons ce point par un exemple artificiel.

### Exemple : Estimateur du total et biais

Reprenons l'exemple concernant les hôpitaux, tirons un échantillon de taille 50 hôpitaux de la population qui en compte 393 (voir Annexe A). Supposons que l'estimateur est  $\hat{T}_0 = N\bar{Y}_s$  qui se justifie si on considère comme modèle de travail  $M : Y_i = \mu + \varepsilon_i$  où les  $\varepsilon_i$  sont indépendantes. Soit  $X$  une variable auxiliaire disponible. Traçons  $y$  en fonction de  $x$ . En regardant la figure 1.3, il semble que  $y = f(x)$  où  $f$  est une fonction linéaire. Le modèle  $M^* : Y_i = \beta x_i + \varepsilon_i$  où les  $\varepsilon_i$  sont indépendantes. Ce modèle est beaucoup plus adapté que le modèle  $M$  aux données. Par rapport au modèle  $M^*$ , l'estimateur du total s'écrit  $\hat{T}_0 = N\bar{Y}_s = N\beta\bar{x}_s$ . Le biais de cet estimateur par rapport au vrai modèle  $M^*$  est :

$$E_{M^*}(\hat{T}_0 - T) = N\beta\bar{x}_s - N\beta\bar{x} = N\beta(\bar{x}_s - \bar{x}),$$

où  $\bar{x}_s$  : moyenne dans l'échantillon et  $\bar{x}$  : moyenne dans la population.

Il apparaît que si  $\bar{x}_s < \bar{x}$  alors  $\hat{T}_0$  est négativement biais, c'est le cas de la figure 1.3 avec

$\bar{x}_s = 102.48 < \bar{x} = 274.6972$ . Alors que dans la figure 1.4,  $\bar{x}_s = 274.78 \approx \bar{x} = 274.6972$

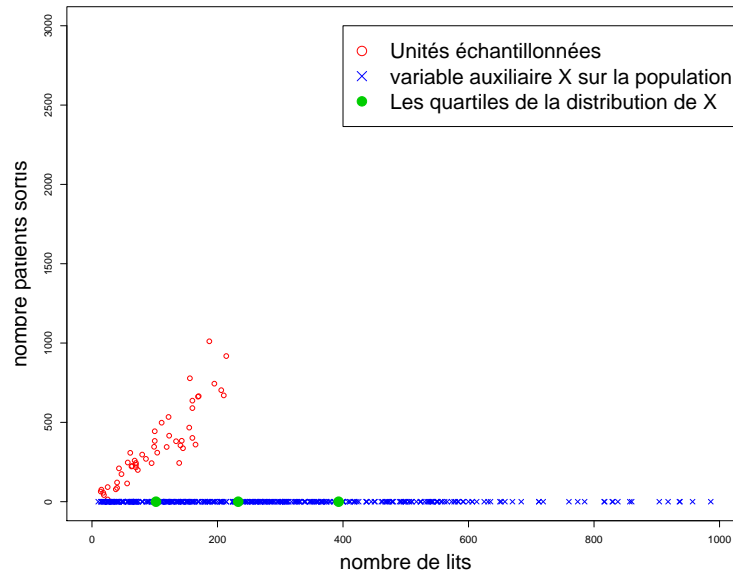


FIG. 1.3 – Échantillon "non équilibré" de la population par rapport à X

donc  $\hat{T}_0$  n'est presque pas biaisé. Cet échantillon est presque équilibré pour la variable auxiliaire  $x$  dans le sens où  $\bar{x}_s \approx \bar{x}$ .

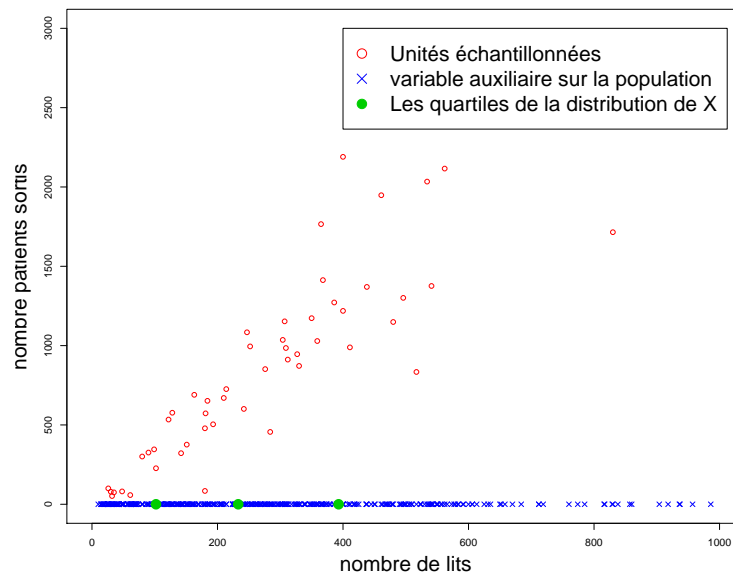


FIG. 1.4 – Échantillon "presque équilibré" de la population par rapport à X

Cet exemple suggère qu'il est possible de se protéger contre le choix d'un mauvais modèle en choisissant des échantillons représentatifs de la population. Ici se protéger signifie conserver le caractère non biaisé de nos prédicteurs. Et dans le cas des modèles plus complexes, la notion d'échantillons représentatifs sera précisée par la définition formelle des échantillons équilibrés.

### 1.3.2 Définition d'un échantillon équilibré et de la stratégie de robustesse au biais

L'échantillon équilibré permet de se protéger contre un modèle de travail éventuellement faux. Ce modèle est qualifié de faux parce qu'il ne représente pas bien ou moins bien les données que le vrai modèle. Définissons formellement un échantillon équilibré par rapport à une variable auxiliaire  $X$  continue.

**Définition 5.** *Un échantillon  $s$  est dit équilibré d'ordre  $J$  et est noté  $s(J)$  s'il vérifie la condition suivante :*

$$\bar{x}_s^{(j)} = \bar{x}^{(j)} \quad \text{pour tout} \quad j = 1, 2, \dots, J$$

$$\text{où } \bar{x}_s^{(j)} = \sum_s x_i^j / n \text{ et } \bar{x}^{(j)} = \sum_{i=1}^N x_i^j / N.$$

La stratégie consiste en la combinaison du choix de l'estimateur avec un type d'échantillonnage donné. La stratégie de robustesse du biais est le cas particulier où le caractère non biaisé du prédicteur est conservé. Dans les exemples de la prochaine section, il sera montré grâce aux exemples que l'estimateur du total combiné à un échantillon équilibré est une stratégie de robustesse du biais, ainsi que l'estimateur du ratio combiné avec un échantillon équilibré.

Remarquons que l'estimateur du total par le ratio,  $\hat{T}(0, 1 : x) = N\hat{y}_s\bar{x}/\bar{x}_s$ , se réduit à l'estimateur du total par la moyenne  $\hat{T}(1 : 1) = N\hat{y}_s$  lorsque l'échantillon est équilibré d'ordre 1  $s(1)$  c'est à dire que  $\bar{x}_s = \bar{x}$ . Ce phénomène très intéressant peut être généralisé pour les meilleurs prédicteurs linéaires non biaisés de modèles plus complexes. Ainsi des prédicteurs complexes peuvent être réduits à l'estimateur du total par la moyenne. Le théorème qui suit généralise le phénomène.

**Théorème 2** (Royall et Herson 1973a). *Si  $s = s(J)$  alors*  
 $\hat{T}(\delta_0, \dots, \delta_j, \dots, \delta_J : x^j) = \hat{T}(1 : 1)$  *à condition que  $\delta_j = 1$  et  $j \in \{1, \dots, J\}$ .*

Pour la démonstration voir Royall et Hersen (1973a).

Ce théorème précise que le meilleur prédicteur linéaire non biaisé d'un échantillon équilibré se réduit à l'estimateur de la moyenne si le modèle a une variance de  $Y$  proportionnelle à  $x$  et que le coefficient de  $x^j$  n'est pas nul dans le modèle. L'estimateur du total par le ratio respecte ces conditions donc peut être réduit à l'estimateur du total par la moyenne. De ce théorème découle le corollaire suivant :

**Corollaire 2.** *L'estimateur  $\hat{T}(\delta_0, \dots, \delta_j, \dots, \delta_J : x^j)$  avec  $\delta_j = 1$  combiné à un échantillon  $s = s(J)$  est une stratégie de robustesse au biais dans le cadre des modèles polynômiaux.*

L'intérêt de ce théorème est évident. Souvent pour représenter les données de façon très satisfaisante, il faut utiliser un polynôme de degré élevé. Dans certain cas il sera donc possible de substituer le meilleur prédicteur linéaire non biaisé souvent complexe qui découle de ce modèle par l'estimateur du total par la moyenne. Cette simplification substantielle du prédicteur n'affecte pas les estimations ponctuelles.

### Exemple : Estimateur du total par la régression linéaire

L'estimateur de la régression linéaire est

$$\hat{T}_{RL} = N(\bar{Y}_s + \hat{\beta}_1(\bar{x} - \bar{x}_s)), \quad \text{où} \quad \hat{\beta}_1 = \sum_s (y_i - \bar{y}_s)(x_i - \bar{x}_s) / \left( \sum_s x_i^2 - n\bar{x}_s^2 \right).$$

Cet estimateur correspond au meilleur prédicteur linéaire non biaisé sous  $M(1, 1 : 1)$  avec une variance de  $N^2(1-f)\sigma^2[1+(\bar{x}_s-\bar{x})^2/\{(1-f)c_s\}]/n$ , où  $c_s = \sum_s (x_i - \bar{x}_s)^2/n$ . D'après le théorème (2), si  $s = s(1)$  c'est à dire  $\bar{x}_s = \bar{x}$  alors  $\hat{T}_{RL}$  se réduit à l'estimateur du total  $\hat{T}(1 : 1)$ . Et il est non biaisé sous le modèle général (1.1) si  $s = s(J)$ . Notez aussi que la variance se réduit à la variance de l'estimateur du total par la moyenne si  $s = s(J)$ .

Alors on peut se demander s'il est utile de se préoccuper des prédicteurs complexes, comme celui de la régression ou plus généralement  $\hat{T}(\delta_0, \dots, \delta_j, \dots, \delta_J : x^j)$ , sous un échantillon équilibré  $s = s(J)$  quand il suffit juste de les remplacer par l'estimateur du total  $\hat{T}(1 : 1)$ . La réponse est oui du fait que :

- il faut souvent souligner que les estimateurs de variance ne seront pas les mêmes puisque les résidus sous les différents modèles ne sont pas identiques même avec des échantillons équilibrés.
- il est difficile de choisir des échantillons équilibrés surtout à des degrés élevés. Dans cette situation, les estimations vont être différemment affecté par le degré

d'équilibrage de l'échantillon. Généralement l'estimateur du total par le ratio est préférable à l'estimateur par la moyenne sous un échantillon presque équilibré (voir p. 57 de Valliant, Dorfman et Royall (2000)).

### 1.3.3 Quelques estimateurs sous échantillons équilibrés

#### Exemple : Estimateur du total par la moyenne sous échantillon équilibré

L'estimateur du total coïncide avec avec le meilleur prédicteur linéaire non biaisé sous le modèle  $M(1 : 1)$ . Cet estimateur s'écrit  $\hat{T}(1 : 1) = N\hat{Y}_s$ . Si le vrai modèle  $M^*$  est le modèle général  $Y_i = \sum_{j=0}^J \delta_j \beta_j x_i^j + \varepsilon_i \gamma_i^{1/2}$  alors le biais de l'estimateur est :

$$\begin{aligned} E_{M^*}[\hat{T}(1 : 1) - T] &= E_{M^*}[N\bar{Y}_s - N\bar{Y}] \\ &= N * E_{M^*}[\sum_{j=1}^J \delta_j \beta_j \bar{x}_s^{(j)} - \sum_{j=1}^J \delta_j \beta_j \bar{x}^{(j)}] \\ &= N \sum_{j=1}^J \delta_j \beta_j E_{M^*}[\bar{x}_s^{(j)} - \bar{x}^{(j)}] \\ &= N \sum_{j=1}^J \delta_j \beta_j [\bar{x}_s^{(j)} - \bar{x}^{(j)}], \end{aligned}$$

où  $\bar{x}_s^{(j)} = \sum_s x_s^j / n$  et  $\bar{x}^{(j)} = \sum_{i=1}^N x_i^j / N$ .

Le biais est donc nul si l'échantillon est équilibré jusqu'au degré  $J$  c'est à dire  $s = s(J)$ . En d'autres termes, l'estimateur du total est non biaisé pour tous les modèles polynômiaux de la famille de (1.1) en autant que l'échantillon soit équilibré jusqu'au degré du polynôme.

Par exemple, si le vrai modèle est  $M^* = M(1, 1 : 1)$  alors le biais de l'estimateur du total  $\hat{T}(1 : 1)$  est :

$$E_{M^*}[\hat{T}(1 : 1) - T] = N\beta_1(\bar{x}_s - \bar{x}).$$

Sous le vrai modèle  $M^* = M(1, 1 : 1)$ , l'échantillon équilibré protège contre le biais quand on utilise l'estimateur du total par la moyenne.

**Exemple : Estimateur du total sous le modèle du ratio sous échantillon équilibré**

Supposons que le vrai modèle est  $M^* = M(1, 1 : x)$  c'est à dire  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ . Le modèle de travail est  $M(0, 1 : x)$  donc le meilleur prédicteur linéaire est l'estimateur du ratio  $\hat{T}_R = \hat{T}(0, 1 : x) = N\bar{Y}_s \bar{x} / \bar{x}_s$ . Le biais dans ce cas est :

$$\begin{aligned}
 E_{M^*}[\hat{T}(0, 1 : x) - T] &= E_{M^*}[N\bar{Y}_s \bar{x} / \bar{x}_s - N\bar{Y}] \\
 &= N * E_{M^*}[\bar{Y}_s \bar{x} / \bar{x}_s] - N\bar{Y} \\
 &= N(\beta_0 + \beta_1 \bar{x}_s) \bar{x} / \bar{x}_s - N(\beta_0 + \beta_1 \bar{x}) \\
 &= N(\beta_0 \bar{x} / \bar{x}_s + \beta_1 \bar{x}) - N(\beta_0 + \beta_1 \bar{x}_s) \\
 &= N(\beta_0 \bar{x} / \bar{x}_s + \beta_1 \bar{x} - \beta_0 - \beta_1 \bar{x}_s) \\
 &= N(\beta_0 \bar{x} / \bar{x}_s - \beta_0) \\
 &= N\beta_0(\bar{x} - \bar{x}_s) / \bar{x}_s
 \end{aligned}$$

Le biais de l'estimateur du ratio sous le vrai modèle  $M^*$  est :

$$E_{M^*}[\hat{T}(0, 1 : x) - T] = N\beta_0(\bar{x} - \bar{x}_s) / \bar{x}_s$$

Constatons que de la même façon que précédemment, si l'échantillon est équilibré c'est à dire si  $\bar{x}_s = \bar{x}$  alors l'estimateur du ratio est non biaisé pour le vrai modèle  $M(1, 1 : x)$ . Il est possible de généraliser ce résultat. Pour cela considérons que le vrai modèle est le modèle polynômial général (1.1), il vient que :

$$\begin{aligned}
 E_{M^*}[\hat{T}(0, 1 : x) - T] &= E_{M^*}[N\bar{Y}_s \bar{x} / \bar{x}_s - N\bar{Y}] \\
 &= N[\sum_{j=1}^J \delta_j \beta_j \bar{x}_s^{(j)} \bar{x} / \bar{x}_s - \sum_{j=1}^J \delta_j \beta_j \bar{x}^{(j)}] \\
 &= N \sum_{j=1}^J \delta_j \beta_j [\bar{x}_s^{(j)} \bar{x} / \bar{x}_s - \bar{x}^{(j)}]
 \end{aligned}$$

Le biais de l'estimateur du ratio sous ce modèle s'écrit alors

$$E_{M^*}[\hat{T}(0, 1 : x) - T] = N\bar{x} \sum_{j=0}^J \delta_j \beta_j [\bar{x}_s^{(j)} / \bar{x}_s - \bar{x}^{(j)} / \bar{x}]$$

Si  $\bar{x}_s^{(j)} = \bar{x}^{(j)}$  quelque soit  $j \in \{1, \dots, J\}$  alors ce biais est nul. En conclusion, sous le vrai modèle général (1.1), l'estimateur du total par le ratio de la même façon que l'estimateur du total par la moyenne est non biaisé si l'échantillon est équilibré jusqu'au degré  $J$ .

# Chapitre 2

## Modèles avec variables qualitatives

La question qui peut se poser est de savoir : comment ajuster des modèles de prédiction contenant des variables auxiliaires qualitatives ? Cette question est d'autant plus pertinente que dans beaucoup d'enquêtes, les variables qualitatives sont d'une importance capitale. On peut penser aux variables sexe, race et âge dans les enquêtes sur la population active par exemple. Dans ses sondages, on est intéressé à connaître l'âge, le sexe, la race de chaque répondant dans le but d'estimer des totaux, des moyennes (exemple de revenu) dans chacun des groupes formés par les différentes classes des variables qualitatives. Les variables quantitatives dans de tels modèles sont appelées covariables, elles peuvent être présentes ou absentes du modèle à variables qualitatives. En présence de variables qualitatives, il arrive souvent que les conditions du théorème général de la prédiction (1) ne soient plus valides en particulier le fait que  $X'_s V_{ss}^{-1} X_s$  doit être inversible. La technique des inverses généralisés permet de résoudre ce problème. Nous conserverons autant que possible les notations du chapitre précédent.

### 2.1 Introduction

Les modèles linéaires avec variables qualitatives sont bien connus en régression et en planification d'expérience. Nous utiliserons le vocabulaire des plans d'expérience ainsi nous parlerons de facteur pour dire variable qualitative et les catégories dans lesquelles chaque facteur est divisé sont appelés niveaux. Searle (1971, 1987) expose de façon plus précise et détaillée les mathématiques qui sous-tendent les modèles linéaires. L'exemple simple suivant va permettre d'introduire les notions de base.

### 2.1.1 Exemple simple

Considérons un exemple tiré de Searle (1971, sec. 4.4). Supposons que nous avons un seul facteur qui est le plus haut degré d'éducation atteint par une personne. Ce facteur possède trois niveaux qui sont (a) Etude secondaire ou moins (b) Etude collégiale (c) Etude universitaire. La variable d'intérêt  $Y$  est le revenu personnel. L'étude essaye d'expliquer le revenu d'une personne en fonction de son degré de scolarisation. Pour cela on peut ajuster un modèle de régression avec le revenu personnel  $Y$  comme variable dépendante et le plus haut degré de scolarisation comme variable explicative. Créons trois variables "dummy" ("artificielles") qui prennent les valeurs 0 ou 1 :

$$x_{1i} = \begin{cases} 1 & \text{si la personne s'est arrêtée à l'école secondaire ou avant} \\ 0 & \text{sinon} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{si la personne s'est arrêtée au collège} \\ 0 & \text{sinon} \end{cases}$$

$$x_{3i} = \begin{cases} 1 & \text{si la personne a un diplôme universitaire} \\ 0 & \text{sinon} \end{cases}$$

Ces trois variables définissent des espaces disjoints. Chaque personne se retrouve dans une seule des trois catégories de niveau d'étude. Ainsi pour toute personne  $i$ , une des variables  $x_{1i}$ ,  $x_{2i}$  et  $x_{3i}$  sera égale à 1 et les autres vaudront 0. Il est donc possible comme pour les variables quantitatives d'ajuster un modèle :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \quad (2.1)$$

Pour simplifier, supposons que les erreurs  $\varepsilon_i$  sont de moyenne 0 et de variance  $\sigma^2$ . Le fait de considérer une variance commune est raisonnable car souvent en pratique il y a peu d'évidence d'hétérogénéité.

Le modèle se complique, on part d'un facteur pour aboutir à un modèle à trois variables "artificielles". Cependant il est au moins possible si le nombre d'observations est suffisant d'ajuster un modèle à variables qualitatives qu'avec les variables quantitatives.

Prenons un échantillon de taille  $n = 6$  personnes pour faire une petite application numérique. Choisissons trois personnes s'étant arrêtés à l'école secondaire ou avant, deux personnes ayant arrêté au collège et une seule personne a fait des études universitaires. Pour faciliter la compréhension et l'écriture mathématique, la notation suivante sera adoptée.  $Y_{ij}$  désignera le revenu de la  $j^{\text{ème}}$  personne ayant arrêté au  $i^{\text{ème}}$  degré de scolarisation. La forme matricielle du modèle est  $Y_s = X_s \beta + \varepsilon_s$ , où  $s$  désigne l'échantillon

tiré. L'équation s'écrit alors :

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{31} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \end{pmatrix}$$

Comme  $E(\varepsilon_{ij}) = 0$  alors la valeur espérée pour toute personne dans le  $i$ ème niveau d'éducation est  $E_M(Y_{ij}) = \beta_0 + \beta_i = \mu_i$ . Comme dans le chapitre précédent, la méthode des moindres carrés permet d'estimer le vecteur des paramètres  $\beta$  en résolvant l'équation normale

$$X'_s X_s \beta = X'_s Y_s. \quad (2.2)$$

Compte tenu de notre paramétrisation, la matrice  $X_s$  n'est pas de plein rang car la première colonne est la somme des 3 autres colonnes. Cela a pour conséquence l'existence d'une infinité de solution  $\hat{\beta}$  à l'équation normale.

Par exemple pour  $Y'_s = (16, 10, 19, 11, 13, 27)$ , l'équation normale s'écrit

$$\begin{pmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 96 \\ 45 \\ 24 \\ 27 \end{pmatrix}$$

$X'_s X_s$  est singulière car la première colonne est une combinaison linéaire des trois autres. L'équation n'aura donc pas une solution unique.  $\beta^0 = (16, -1, -4, 11)'$  est une solution à l'équation et d'une façon générale tous les  $\beta^d = \beta^0 + (d, -d, -d, -d)'$  sont aussi solution de l'équation normale.

Résolution de l'équation normale, le produit matriciel donne :

$$\begin{cases} 6\beta_0 + 3\beta_1 + 2\beta_2 + \beta_3 = 96 \\ 3\beta_0 + 3\beta_1 = 45 \\ 2\beta_0 + 2\beta_2 = 24 \\ \beta_0 + \beta_3 = 27 \\ 6\beta_0 + 3(15 - \beta_0) + 2(12 - \beta_0) + (27 - \beta_0) = 96 \\ \beta_1 = 15 - \beta_0 \\ \beta_2 = 12 - \beta_0 \\ \beta_3 = 27 - \beta_0 \end{cases}$$

L'équation 1 du système donne  $0\beta_0 = 0$  ce qui est vrai quelque soit la valeur de  $\beta_0$ . Donc les autres composantes de  $\beta$  seront une fonction de  $\beta_0$ . La solution générale est  $\hat{\beta}(\beta_0) = (\beta_0, 15 - \beta_0, 12 - \beta_0, 27 - \beta_0)'$ . Si  $\beta_0 = 16$  on trouve la solution particulière donnée plus haut. il existe une infinité de solution  $\hat{\beta}(\beta_0)$  à l'équation. Il n'est dans ce cas pas possible de parler d'estimateur de  $\beta$  mais seulement d'une solution de l'équation normale.

### 2.1.2 Facteur, niveau et effet

Les termes facteur, niveau et effet sont très familiers en planification d'expériences mais sont beaucoup moins utilisés en théorie de l'échantillonnage. Cela justifie un peu cette section. Ces concepts sont très pratiques dès lors que nous avons à faire à des variables qualitatives dans les modèles de prédiction. D'une façon générale, les variables qualitatives comme le sexe, la race sont appelées des facteurs. Et les différentes positions prises par ce facteur sont appelées niveaux. Prenons l'exemple précédent pour se fixer les idées. Nous avons une seule variable qualitative qui est le plus haut degré d'éducation atteint. Ce facteur a trois niveaux qui sont (a) étude secondaire ou moins, (b) étude collégiale et (c) étude universitaire. L'effet peut être vu comme une quantification de l'action des facteurs sur la variable réponse  $Y$  selon leurs niveaux. Cette façon de voir est celle qu'on rencontre en Analyse de la Variance. En d'autres termes, l'effet montre l'intensité de l'action de chaque niveau des facteurs sur la variable d'intérêt  $Y$ .

Dans la théorie de l'échantillonnage, l'intérêt est tout autre. Il ne s'agit plus d'isoler les effets des facteurs sur la variable réponse. Il s'agit plutôt de sélectionner les facteurs et les niveaux qui permettent la meilleure prédiction possible de la variable  $Y$ . En continuant sur le même exemple, le modèle (2.1) peut s'écrire en terme d'effet :

$$Y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij}, \quad \text{où } \varepsilon_{ij} \sim (0, \sigma^2) \quad i = 1, 2, 3. \quad (2.3)$$

Dans cette représentation, les paramètres  $\beta_1, \beta_2$  et  $\beta_3$  sont les effets des 3 niveaux du facteur "plus haut degré d'éducation".

Il y a deux grandes familles d'effets, fixe et aléatoire. Les effets fixes sont associés aux facteurs avec un nombre fini (restreint) de niveaux et où les niveaux sont fixés lors de la planification de l'enquête. L'autre famille des effets aléatoires est associée aux facteurs avec un très grand nombre de niveaux voir infini. Dans cette situation seulement un nombre fini voir très limité de tous les niveaux possibles est observé. Un cas très fréquent dans les enquêtes sur la population finie est l'effet de l'enquêteur. Cet effet est aléatoire car seulement une portion de toutes les possibilités est considérée. En prenant le facteur enquêteur comme aléatoire, on réduit le biais dû à la non prise en compte de tous les niveaux du facteur. Le facteur enquêteur est donc souvent pris comme aléatoire avec une moyenne de 0 et une variance inconnue (voir Biemer and Stokes, 1985, 1989, 1991). La variance du modèle devient donc la variance des erreurs plus celle du facteur aléatoire enquêteur. Les modèles ne contenant que des facteurs fixes sont appelés modèles à effets fixes et ceux qui n'ont que des facteurs aléatoires sont les modèles à effets aléatoires. Et on devine aisément, les modèles à effets mixtes sont les modèles avec des facteurs fixes et aléatoires.

En théorie de l'échantillonnage souvent les cellules formées par les niveaux des facteurs (tableaux de contingentement) n'ont pas le même nombre d'individus échantillonnés. Le fait d'avoir des tailles égales dans toutes les cellules permet de simplifier de beaucoup

l'estimation des paramètres du modèle. Les méthodes présentées ici sont pour le cas le plus général c'est à dire avec des tailles de cellules différentes.

## 2.2 Inverse généralisé

L'étude de l'inverse généralisé est motivé par le fait que l'équation normale  $X'_s X_s \beta = X'_s Y_s$  n'a pas de solution unique.  $X'_s X_s$  n'étant pas inversible car  $X_s$  n'est pas de plein rang, on est amené à trouver une méthode plus globale pour contourner cet obstacle. Comme au chapitre précédent, nous avons le modèle linéaire général :

$$\begin{aligned} E_M(Y) &= X\beta \\ \text{var}_M(Y) &= V, \end{aligned}$$

où  $X$  est une matrice  $N \times p$ ,  $\beta$  est un vecteur  $p \times 1$  de paramètres inconnus, et  $V$  est une matrice de variance-covariance définie positive.

Pour estimer  $\beta$ , il faut inverser la matrice  $X'_s V_{ss}^{-1} X_s$  qui n'est pas de plein rang. Pour ce faire, nous allons utiliser la méthode des inverses généralisés. Pour une documentation plus complète vous pouvez consulter Searle (1971).

**Définition 6.** *Un inverse généralisé (ou g-inverse) de la matrice  $A$  est défini comme étant toute matrice  $G$  telle que  $AGA = A$ .*

La matrice  $A$  peut être non symétrique ou ne pas être carrée. Ici la matrice  $A_s = X'_s A X_s$  qui nous intéresse est carrée et symétrique. Pour obtenir une solution de l'équation normale (2.2), nous utiliserons les lemmes suivants :

**Lemme 1.** *Si  $G$  est un g-inverse de  $A_s = X'_s V_{ss}^{-1} X_s$  alors  $\beta^0 = G X'_s V_{ss}^{-1} Y_s$  est une solution de l'équation normale  $X'_s V_{ss}^{-1} X_s \hat{\beta} = X'_s V_{ss}^{-1} Y_s$ .*

La preuve de ce lemme nécessite la considération des deux lemmes qui suivent :

**Lemme 2.** *Soient  $X$ ,  $P$  et  $Q$  des matrices de réels de dimensions respectives  $n \times p$ ,  $q \times p$  et  $q \times p$  alors :*

- (i)  $X'X = 0$  implique  $X = 0$
- (ii)  $PX'X = QX'X$  implique  $PX' = QX'$

**Preuve du lemme 2 :**

(i) Si  $X'X = 0$  alors on a que la diagonale est nulle c'est à dire que

$$\text{diag}(X'X) = \begin{pmatrix} \sum_{i=1}^p x_{1i}^2 \\ \vdots \\ \sum_{i=1}^p x_{ni}^2 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

La première équation donne  $\sum_{i=1}^p x_{1i}^2 = 0 \Leftrightarrow x_{1i} = 0 \forall i$ . L'ensemble des  $n$  équations donne  $x_{1i} = \dots = x_{ni} = 0 \quad \forall i \in \{1, \dots, n\}$ . Donc  $X = 0$ .

(ii)

$$\begin{aligned} PX'X &= QX'X \\ PX'X - QX'X &= 0 \\ (PX'X - QX'X)(P - Q)' &= 0 \\ (PX' - QX')X(P - Q)' &= 0 \\ (PX' - QX')(PX' - QX')' &= 0 \end{aligned}$$

D'après (i)  $(PX' - QX') = 0$  c'est à dire  $PX' = QX'$ .

✚

**Lemme 3.** Si  $G$  est un  $g$ -inverse de  $X'X$  alors :

(i)  $G'$  est aussi un  $g$ -inverse de  $X'X$ .

(ii)  $XGX'X = X$  ce qui veut dire que  $GX'$  est un  $g$ -inverse de  $X$ .

(iii)  $XGX'$  est invariant par rapport à  $G$  ce qui veut dire que  $XGX' = XFX'$  pour tout autre  $g$ -inverse  $F$ .

(iv)  $XGX'$  est symétrique que  $G$  le soit ou pas.

**Preuve du lemme 3 :**

(i)  $G$  est un  $g$ -inverse de  $X'X$  donc  $X'XGX'X = X'X$ . Si on transpose on a :  $X'XG'X'X = X'X$  donc  $G'$  est un  $g$ -inverse de  $X'X$ .

(ii) Posons  $P = X'XG'$  et  $Q = I$

La partie (ii) du lemme 2 donne :  $X'XG'X'X = X'X$  implique  $X'XG'X' = X'$ . On transpose et on obtient  $XGX'X = X$ . Donc  $GX'$  est un  $g$ -inverse pour  $X$ .

(iii) Supposons que  $F$  est un autre  $g$ -inverse différent de  $G$ . D'après (ii)  $FX'$  et  $GX'$  sont des  $g$ -inverses de  $X$ . D'où  $XGX'X = X$  et  $XFX'X = X$  les deux égalités donnent  $XGX'X = XFX'X$ . D'après la partie (ii) du lemme 2  $XGX'X = XFX'X$  implique  $XGX' = XFX'$  (où  $P = XG$  et  $Q = XF$ ). Donc  $XGX'$  est invariant par rapport à  $G$ .

(iv) D'après (i),  $G$  et  $G'$  sont tous deux des g-inverses de  $X'X$ . (iii) de ce lemme donne  $XGX' = XG'X'$ , or  $XG'X' = (XGX')'$ . Donc  $XGX'$  est symétrique.

✠

### Preuve du lemme 1 :

Tout d'abord remarquons que  $A_s = A'_s$  car  $A'_s$  est symétrique, posons  $Z_s = V_{ss}^{-1/2}X_s$  et comme  $G$  et  $G'$  sont g-inverse de  $A_s$  donc  $Z_sGZ_s = Z_sG'Z_s$ . Alors :

$$\begin{aligned} X'_sV_{ss}^{-1}X_s\beta^0 &= X'_sV_{ss}^{-1}X_sGX'_sV_{ss}^{-1}Y_s \\ &= A_sGZ'_sV_{ss}^{-1/2}Y_s \\ &= (Z_sG'A'_s)'V_{ss}^{-1/2}Y_s \\ &= (Z_sGA_s)'V_{ss}^{-1/2}Y_s \end{aligned}$$

Or  $A_s = Z'_sZ_s$  et de plus le lemme 3 (ii) montre que  $Z'_sGZ_sZ'_s = Z'_s$ . D'où

$$\begin{aligned} X'_sV_{ss}^{-1}X_s\beta^0 &= (Z_sGZ'_sZ_s)'V_{ss}^{-1/2}Y_s \\ &= (Z_s)'V_{ss}^{-1/2}Y_s \\ &= Z'_sV_{ss}^{-1/2}Y_s \\ &= X'_sV_{ss}^{-1}Y_s \end{aligned}$$

$\beta^0 = GX'_sV_{ss}^{-1}Y_s$  est bien une solution de l'équation normale.

✠

Sous le modèle linéaire général (1.2), la variance de  $\beta^0$  : est

$$var_M(\beta^0) = var_M(GX'_sV_{ss}^{-1}Y_s) = GX'_sV_{ss}^{-1}V_{ss}V_{ss}^{-1}X_sG' = GX'_sV_{ss}^{-1}X_sG' = GA_sG'$$

Si la matrice  $A_s$  est de plein rang alors  $var_M(\beta^0) = G$ . Il est possible de construire un g-inverse de  $A_s$  avec les propriétés :

$$GA_sG' = G \tag{2.4}$$

$$GG' \quad \text{et} \quad G'G \quad \text{sont} \quad \text{symétriques} \tag{2.5}$$

Un g-inverse satisfaisant les propriétés 2.4 et 2.5 est appelé un inverse Moore-Penrose et est unique (Gentle, 1998, sec. 2.1.9). Les g-inverse des matrices symétriques produits

par la méthode "décomposition en valeur singulière" (SVD) (voir annexe B) sont des inverses Moore-Penrose et sont eux même symétriques. Dans la suite du chapitre, certains inverses généralisés seront considérés symétriques et la propriété 2.4 vraie. Cela ne réduit en rien la généralité du résultat mais conduit parfois à des expressions plus simples.

Même si le  $g$ -inverse n'est pas unique donc la solution  $\beta^0$  non plus, la section suivante prouve que certaines fonctions de  $\beta^0$  sont quant à elles uniques. Une famille de ces fonctions conduit à des estimateurs et est connue sous le nom de "fonctions estimables" (Searle 1971, sec. 5.4).

## 2.3 Estimation d'une combinaison linéaire des $Y_i$

Comme dans le chapitre précédent ce qui nous intéresse ici c'est d'estimer une combinaison linéaire des  $Y_i$ , notée  $\theta = \gamma'Y$ . Nous considérons le modèle linéaire général (1.2) :

$$\begin{aligned} E_M(Y) &= X\beta \\ \text{var}_M(Y) &= V, \end{aligned}$$

L'approche est la même qu'au chapitre 1 mais du fait que  $X'_s X_s$  n'est pas de plein rang, la solution  $\hat{\beta}$  de l'équation normale n'est pas unique. Dans ces conditions peut-on trouver une estimation de  $\beta$ ? Et mieux encore comment trouver l'estimateur optimal si on a une infinité de solution pour  $\beta$ ? Nous rappelons que l'indice  $r$  désigne les unités non échantillonnées et l'indice  $s$  fait référence à l'échantillon.

### 2.3.1 Cas général

Pour répondre aux questions précédentes, nous avons besoins d'un lemme supplémentaire.

**Lemme 4.** *Si  $G$  est un  $g$ -inverse de  $A_s = X'_s V_{ss}^{-1} X_s$  alors :*

(i)  $X_r G X'_s$  est invariant au choix de  $G$ .

(ii)  $X_r G A_s = X_r$

**Preuve du lemme 4 :**

(i) Soit  $F$  un autre  $g$ -inverse de  $A_s$ . Soit  $q$  le rang de  $X$ , écrivons  $X = [X_1 \ X_2]$ , où  $X_1$  est la matrice de plein rang de dimension  $N \times q$ ,  $X_2$  est la matrice de dimension  $N \times (p-q)$ . Comme les colonnes de  $X_2$  sont des combinaisons linéaires de celles de  $X_1$  alors  $X_2 =$

$X_1K$  où  $K$  est une certaine matrice  $q \times (p-q)$ . Nous pouvons écrire  $X = X_1K^*$ , où  $K^* = [I_q \ K]$  et  $I_q$  est la matrice identité  $q \times q$ . Il est possible d'écrire  $X$  en terme d'unités échantillonnées et non échantillonnées  $X = [X'_s \ X'_r]'$  où  $X_s = [X_{1s} \ X_{2s}] = X_{1s}K^*$  où  $X_{1s}$  est une matrice de dimension  $n \times q$  et  $X_r = X_{1r}K^*$ . Comme  $F$  et  $G$  sont des  $g$ -inverses de  $A_s$ , le lemme 3 (iii) implique que  $V_{ss}^{-1/2}X_sGX'_sV_{ss}^{-1/2} = V_{ss}^{-1/2}X_sFX'_sV_{ss}^{-1/2}$ . Ce qui donne

$$V_{ss}^{-1/2}X_{1s}K^*GX'_sV_{ss}^{-1/2} = V_{ss}^{-1/2}X_{1s}K^*FX'_sV_{ss}^{-1/2} \quad (2.6)$$

$X_{1s}$  est de plein rang et en multipliant l'équation (2.6) par  $X_{1r}A_{1s}^{-1}X'_{1s}V_{ss}^{-1/2}$  avec  $A_{1s} = X'_{1s}V_{ss}^{-1}X_{1s}$  sur la gauche et par  $V_{ss}^{1/2}$  sur la droite, nous obtenons  $X_rGX'_s = X_rFX'_s$ .

(ii) Par le lemme 3,

$$V_{ss}^{-1/2}X_sGA_s = V_{ss}^{-1/2}X_s$$

En utilisant la même décomposition de  $X$  qu'en (i), nous obtenons :

$$V_{ss}^{-1/2}X_{1s}K^*GA_s = V_{ss}^{-1/2}X_{1s}K^*.$$

Multiplions chaque côté de l'équation par  $X_{1r}A_{1s}^{-1}X'_{1s}V_{ss}^{-1/2}$  ce qui donne  $X_rGA_s = X_r$ .

✠

Dans le cas du chapitre 1, il a été montré que si  $X'_sV_{ss}^{-1}X_s$  est inversible alors le meilleur estimateur linéaire non biaisé est  $\hat{\theta}_{opt} = \gamma'_sY_s + \gamma'_r[X_r\hat{\beta} + V_{rs}V_{ss}^{-1}(Y_s - X_s\hat{\beta})]$  (voir le théorème général 1). Comme  $X'_sV_{ss}^{-1}X_s$  n'est pas inversible, il existe une infinité de solution de l'équation normale. On peut alors se demander si on obtient un  $\hat{\theta}_{opt}$  différent pour chaque solution  $\hat{\beta}_i$ . Heureusement que le théorème suivant répond à la question par la négative.

**Théorème 3.** *L'estimateur*

$$\hat{\theta}_{opt} = \gamma'_sY_s + \gamma'_r[X_r\beta^0 + V_{rs}V_{ss}^{-1}(Y_s - X_s\beta^0)],$$

où  $\hat{\beta}^0 = GX'_sV_{ss}^{-1}Y_s$  et  $G$  est un  $g$ -inverse de  $X'_sV_{ss}^{-1}X_s$ , est invariant au choix de  $G$ .

**Preuve du théorème 3 :**

$$\begin{aligned} \hat{\theta}_{opt} &= \gamma'_sY_s + \gamma'_r[X_r\beta^0 + V_{rs}V_{ss}^{-1}(Y_s - X_s\beta^0)] \\ \hat{\theta}_{opt} &= \gamma'_sY_s + \gamma'_r[X_rGX'_s + V_{rs} - V_{rs}V_{ss}^{-1}X_sGX'_s]V_{ss}^{-1}Y_s \end{aligned}$$

D'après le lemme 4,  $X_rGX'_s$  est invariant au choix de  $G$ . De même le lemme 3 (iii) montre que  $V_{ss}^{-1/2}X_sGX'_sV_{ss}^{-1/2}$  est invariant au choix de  $G$ . Donc  $\hat{\theta}_{opt}$  est invariant comme somme de deux quantités invariantes.



Nous avons trouvé un estimateur de  $\theta$ , il reste à montrer que c'est le meilleur estimateur linéaire  $\theta$ . Pour cela il faut montrer que l'estimateur  $\hat{\theta}_{opt}$  est de variance minimale.

**Théorème 4.** *Le meilleur prédicteur linéaire non biaisé de  $\theta = \gamma'Y$  sous le modèle (1.2) est :*

$$\hat{\theta}_{opt} = \gamma'_s Y_s + \gamma'_r [X_r \beta^0 + V_{rs} V_{ss}^{-1} (Y_s - X_s \beta^0)], \quad (2.7)$$

où

$$\beta^0 = G X'_s V_{ss}^{-1} Y_s.$$

La variance de l'erreur de  $\hat{\theta}_{opt}$  est :

$$\text{var}_M(\hat{\theta}_{opt} - \theta) = \gamma'_r (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r + \gamma'_r (X_r - V_{rs} V_{ss}^{-1} X_s) G (X_r - V_{rs} V_{ss}^{-1} X_s)' \gamma_r,$$

où  $G$  est un  $g$ -inverse de  $A_s = X'_s V_{ss}^{-1} X_s$ .

**Preuve du théorème (4) :**

(i) Montrons que  $\hat{\theta}_{opt}$  est sans biais

$$\begin{aligned} \hat{\theta}_{opt} - \theta &= \gamma'_r [(X_r - V_{rs} V_{ss}^{-1} X_s) G X'_s + V_{rs}] V_{ss}^{-1} Y_s - \gamma'_r Y_r \\ E(\hat{\theta}_{opt} - \theta) &= \gamma'_r [X_r G X'_s - V_{rs} V_{ss}^{-1} X_s G X'_s + V_{rs}] V_{ss}^{-1} X_s \beta^0 - \gamma'_r X_r \beta^0 \end{aligned}$$

Le lemme 4 (ii) donne  $X_r G X'_s V_{ss}^{-1} X_s = X_r$  et le lemme 3 (ii) donne  $V_{ss}^{-1/2} X_s G X'_s V_{ss}^{-1/2} X_s = V_{ss}^{-1/2} X_s$  donc  $V_{rs} V_{ss}^{-1} X_s G X'_s V_{ss}^{-1} X_s = V_{rs} V_{ss}^{-1/2} V_{ss}^{-1/2} X_s$

$$\begin{aligned} E(\hat{\theta}_{opt} - \theta) &= \gamma'_r [X_r \beta^0 - V_{rs} V_{ss}^{-1} X_s \beta^0 + V_{rs} V_{ss}^{-1} X_s \beta^0] - \gamma'_r X_r \beta^0 \\ E(\hat{\theta}_{opt} - \theta) &= \gamma'_r [X_r - V_{rs} V_{ss}^{-1} X_s + V_{rs} V_{ss}^{-1} X_s - X_r] \beta^0 \\ E(\hat{\theta}_{opt} - \theta) &= 0. \end{aligned}$$

(ii) Variance minimale

Comme dans le chapitre précédent, la variance d'un estimateur de la forme  $g'_s Y_s$  est :

$$\begin{aligned} E_M(g'_s Y_s - \gamma' Y)^2 &= E_M(a' Y_s - \gamma'_r Y_r)^2 \\ E_M(g'_s Y_s - \gamma' Y)^2 &= \text{var}_M(a' Y_s - \gamma'_r Y_r) + [E_M(a' Y_s - \gamma'_r Y_r)]^2 \\ E_M(g'_s Y_s - \gamma' Y)^2 &= a' V_{ss} a - 2a' V_{sr} \gamma_r + \gamma'_r V_{rr} \gamma_r + [(a' X_s - \gamma'_r X_r) \beta]^2 \end{aligned} \quad (2.8)$$

où  $a = g_s - \lambda_s$ . Le terme entre crochet est nul du fait que l'estimateur est sans biais (voir page 10).

La fonction de Lagrange à minimiser est :

$$L_\lambda(a) = a' V_{ss} a - 2a' V_{sr} \gamma_r + 2(a' X_s - \gamma'_r X_r) \lambda$$

On dérive par rapport à  $a$  et on égalise la dérivée à 0 ce qui donne :

$$\begin{aligned} X_s \lambda &= V_{sr} \gamma_r - V_{ss} a \\ a &= V_{ss}^{-1} (V_{sr} \gamma_r - X_s \lambda) \end{aligned}$$

La démonstration n'est pas très détaillée du fait qu'elle est identique à celle du théorème général **1** dans le chapitre I. La seule différence est que  $A_s$  n'est pas inversible donc on a :

$$\lambda = G(X'_s V_{ss}^{-1} V_{sr} - X'_r) \gamma_r \quad (2.9)$$

où  $G$  est un g-inverse de  $A_s = X'_s V_{ss}^{-1} X_s$ .

On obtient donc la solution pour  $a$  :

$$a_{opt} = V_{ss}^{-1} [V_{sr} + X_s G' (X'_r - X'_s V_{ss}^{-1} V_{sr})] \gamma_r.$$

L'estimateur optimal est donc :

$$\begin{aligned} \hat{\theta}_{opt} &= \gamma'_s Y_s + a'_{opt} Y_s \\ \hat{\theta}_{opt} &= \gamma'_s Y_s + \gamma'_r [V_{rs} + (X_r - V_{rs} V_{ss}^{-1} X_s) G X'_s] V_{ss}^{-1} Y_s \\ \hat{\theta}_{opt} &= \gamma'_s Y_s + \gamma'_r [V_{rs} V_{ss}^{-1} Y_s + X_r G' X'_s V_{ss}^{-1} Y_s - V_{rs} V_{ss}^{-1} X_s G' X'_s V_{ss}^{-1} Y_s]. \end{aligned}$$

Or  $X_r G' X'_s = X_r G X'_s$  grâce au lemme **4** (i)

et  $V_{ss}^{-1/2} X_s G' X'_s V_{ss}^{-1/2} = V_{ss}^{-1/2} X_s G X'_s V_{ss}^{-1/2}$  grâce au lemme **3** (iii)

Ce qui donne :

$$\begin{aligned} \hat{\theta}_{opt} &= \gamma'_s Y_s + \gamma'_r [V_{rs} V_{ss}^{-1} Y_s + X_r G X'_s V_{ss}^{-1} Y_s - V_{rs} V_{ss}^{-1} X_s G X'_s V_{ss}^{-1} Y_s] \\ \hat{\theta}_{opt} &= \gamma'_s Y_s + \gamma'_r [X_r \hat{\beta}^0 + V_{rs} V_{ss}^{-1} (Y_s - X_s \hat{\beta}^0)]. \end{aligned}$$

En substituant  $a_{opt}$  dans l'expression (2.8)

$$\begin{aligned} var_M(\hat{\theta}_{opt} - \theta) &= \gamma'_r (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r + \gamma'_r (B G' X'_s V_{ss}^{-1} V_{sr} + B G' X'_s V_{ss}^{-1} X_s G B' \\ &\quad + V_{rs} V_{ss}^{-1} X_s G B' - 2 B G' X'_s V_{ss}^{-1} V_{sr})' \gamma_r \\ var_M(\hat{\theta}_{opt} - \theta) &= \gamma'_r (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r + \gamma'_r (B G' X'_s V_{ss}^{-1} X_s G B' + V_{rs} V_{ss}^{-1} X_s G B' \\ &\quad - B G X'_s V_{ss}^{-1} V_{sr})' \gamma_r \end{aligned}$$

où  $B = X_r - V_{rs} V_{ss}^{-1} X_s$ .

Or

$$\begin{aligned} B G' X'_s V_{ss}^{-1} X_s &= (X_r - V_{rs} V_{ss}^{-1} X_s) G X'_s V_{ss}^{-1} X_s \\ B G' X'_s V_{ss}^{-1} X_s &= X_r G X'_s V_{ss}^{-1} X_s - V_{rs} V_{ss}^{-1} X_s G X'_s V_{ss}^{-1} X_s \end{aligned}$$

On a :  $X_r G X_s' V_{ss}^{-1} X_s = X_r G A_s = X_r$  d'après les lemmes 3 (i) et 4 (ii).  
Le lemme 3 (ii) donne  $V_{ss}^{-1/2} X_s G A_s = V_{ss}^{-1/2} X_s$ . Donc :

$$\begin{aligned} B G' X_s' V_{ss}^{-1} X_s &= X_r - V_{rs} V_{ss}^{-1/2} V_{ss}^{-1/2} X_s \\ &= X_r - V_{rs} V_{ss}^{-1} X_s \\ &= B. \end{aligned}$$

On obtient alors :

$$\begin{aligned} \text{var}_M(\hat{\theta}_{opt} - \theta) &= \gamma_r'(V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r + \gamma_r'(B G B' + V_{rs} V_{ss}^{-1} X_s G B' \\ &\quad - B G' X_s' V_{ss}^{-1} V_{sr})' \gamma_r \\ &= \gamma_r'(V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r + \gamma_r'(B G B') \gamma_r + \gamma_r'(V_{rs} V_{ss}^{-1} X_s G B' \\ &\quad - B G' X_s' V_{ss}^{-1} V_{sr})' \gamma_r. \end{aligned}$$

Or

$$\begin{aligned} V_{rs} V_{ss}^{-1} X_s G B' - B G X_s' V_{ss}^{-1} V_{sr} &= V_{rs} V_{ss}^{-1} X_s G (X_r' - X_s' V_{ss}^{-1} V_{sr}) - (X_r - V_{rs} V_{ss}^{-1} X_s) * \\ &\quad G' X_s' V_{ss}^{-1} V_{sr} \\ &= V_{rs} V_{ss}^{-1} X_s G X_r' - X_r G' X_s' V_{ss}^{-1} V_{sr} \end{aligned}$$

Cette différence est nulle car  $X_s G X_r = X_s G' X_r$ , lemme 4 (i).

✠

### 2.3.2 Estimation du total dans un modèle unifactoriel

Dans un modèle unifactoriel, il y a un seul facteur à plusieurs niveaux. Ce modèle est semblable à un modèle stratifié. Considérons le cas général où nous avons un facteur à  $I$  niveaux et  $N_i$  unités dans la population pour le niveau  $i$ ,  $1 \leq i \leq I$ . Le modèle s'écrit alors :

$$Y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij}, \quad (2.10)$$

où  $\varepsilon_{ij} \sim (0, \sigma^2)$ ,  $i \in \{1, \dots, I\}$  et  $j \in \{1, \dots, N_i\}$ .

Pour constituer l'échantillon, choisissons  $n_i$  unités pour chaque niveau  $i = 1, \dots, I$ . La taille de l'échantillon est donc  $n = \sum_{i=1}^I n_i$ . Pour des raisons de commodité, définissons

certaines variables comme  $s_i$  qui est la portion de l'échantillon issue du niveau  $i$ ,  $Y_{s_i} = \sum_{j \in S_i} Y_{ij}$  et  $Y_s = \sum_{i=1}^I Y_{s_i}$ . L'équation normale s'écrit alors :

$$X'_s X_s \beta^0 = \begin{pmatrix} n & n_1 & n_2 & \dots & n_I \\ n_1 & n_1 & 0 & 0 & 0 \\ n_2 & 0 & n_2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ n_I & 0 & 0 & 0 & n_I \end{pmatrix} \begin{pmatrix} \beta_0^0 \\ \beta_1^0 \\ \beta_2^0 \\ \vdots \\ \beta_I^0 \end{pmatrix} = \begin{pmatrix} Y_s \\ Y_{s_1} \\ Y_{s_2} \\ \vdots \\ Y_{s_I} \end{pmatrix} = X'_s Y_s$$

$X'_s X_s$  est de dimension  $(I+1) \times (I+1)$  et de rang  $I$ . Comme la sous matrice

$$(X'_s X_s)_{I \times I} = \begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & n_I \end{pmatrix} \text{ est inversible alors un g-inverse de } X'_s X_s \text{ est}$$

$G = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \text{diag}(1/n_i) \end{pmatrix}$  où  $\text{diag}(1/n_i)$  est la matrice diagonale de dimension  $I \times I$  avec  $1/n_i$  sur la diagonale.

En effet

$$AGA = \begin{pmatrix} n & n_1 & n_2 & \dots & n_I \\ n_1 & n_1 & 0 & 0 & 0 \\ n_2 & 0 & n_2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ n_I & 0 & 0 & 0 & n_I \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1/n_1 & 0 & 0 & 0 \\ 0 & 0 & 1/n_2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1/n_I \end{pmatrix} \begin{pmatrix} n & n_1 & n_2 & \dots & n_I \\ n_1 & n_1 & 0 & 0 & 0 \\ n_2 & 0 & n_2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ n_I & 0 & 0 & 0 & n_I \end{pmatrix}$$

$$\begin{aligned} AGA &= \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} n & n_1 & n_2 & \dots & n_I \\ n_1 & n_1 & 0 & 0 & 0 \\ n_2 & 0 & n_2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ n_I & 0 & 0 & 0 & n_I \end{pmatrix} \\ &= \begin{pmatrix} n & n_1 & n_2 & \dots & n_I \\ n_1 & n_1 & 0 & 0 & 0 \\ n_2 & 0 & n_2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ n_I & 0 & 0 & 0 & n_I \end{pmatrix} \end{aligned}$$

Un solution de l'équation normale est donc : En effet

$$\begin{aligned} \beta^0 = GX'_s Y_s &= \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1/n_1 & 0 & 0 & 0 \\ 0 & 0 & 1/n_2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1/n_I \end{pmatrix} \begin{pmatrix} Y_s \\ Y_{s_1} \\ Y_{s_2} \\ \vdots \\ Y_{s_I} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ Y_{s_1}/n_i \\ Y_{s_2}/n_i \\ \vdots \\ Y_{s_I}/n_i \end{pmatrix} = \begin{pmatrix} 0 \\ \bar{Y}_{s_1} \\ \bar{Y}_{s_2} \\ \vdots \\ \bar{Y}_{s_I} \end{pmatrix} \end{aligned} \quad (2.11)$$

Constatons que la solution serait la même si nous avions pris l'inverse de la sous matrice  $(X'_s X_s \beta^0)_I$ . D'une façon générale, la méthode consiste à :

- Déterminer le nombre de paramètres estimables dans le modèle c'est à dire le rang de  $X'_s X_s$  noté  $r(X'_s X_s)$ .
- Calculer la différence entre l'ordre (nombre de colonnes) de  $X'_s X_s$  et son rang :  $p - r(X'_s X_s)$ .
- Fixer  $p - r(X'_s X_s)$  paramètres à 0 et réduire  $X_s$  en excluant les colonnes correspondant aux paramètres fixés à 0. Cela de telle façon que la matrice réduite  $X_{S(R)}$  soit de plein rang. Il faut donc éliminer les colonnes qui sont combinaison linéaire des autres.
- Estimer les paramètres avec la relation  $\hat{\beta}(R) = [X'_{S(R)} X_{S(R)}]^{-1} X'_{S(R)} Y_s$

Cette méthode que nous venons de décrire est équivalente à calculer un g-inverse. L'estimateur du total correspondant est :

$$\begin{aligned} \hat{T} &= 1'_s Y_s + 1'_r X_r \beta^0 \\ \hat{T} &= \sum_{i=1}^I n_i \bar{Y}_{si} + \sum_{i=1}^I (N_i - n_i) \bar{Y}_{si} \\ \hat{T} &= \sum_{i=1}^I N_i \bar{Y}_{si} \end{aligned}$$

En accord avec ce qui a été dit précédemment, on obtient le même estimateur que dans le cas stratifié. Pour que cet estimateur soit rigoureusement valide, on suppose qu'on a des observations dans toutes les modalités du facteur.

Une autre façon de trouver une solution  $\beta^0$  est de rajouter des contraintes. Une contrainte

souvent utilisée est  $\sum_{i=1}^I n_i \beta_i = 0$  dans ce cas la solution est :

$$\beta^0 = \begin{pmatrix} \bar{Y}_s \\ \bar{Y}_{s_1} - \bar{Y}_s \\ \bar{Y}_{s_2} - \bar{Y}_s \\ \vdots \\ \bar{Y}_{s_I} - \bar{Y}_s \end{pmatrix} \quad (2.12)$$

L'estimateur du total correspondant est :

$$\begin{aligned} \hat{T} &= \mathbf{1}'_s Y_s + \mathbf{1}'_r X_r \beta^0 \\ \hat{T} &= \sum_{i=1}^I n_i \bar{Y}_{si} + (N - n) \bar{Y}_s + \sum_{i=1}^I (N_i - n_i) (\bar{Y}_{si} - \bar{Y}_s) \\ \hat{T} &= \sum_{i=1}^I N_i \bar{Y}_{si} \end{aligned}$$

On trouve le même estimateur, ce qui est normal car nous avons montré par le théorème 3 que l'estimateur est invariant au choix de  $G$  c'est à dire au choix d'une solution particulière de  $\beta$ .

Calculons la variance des erreurs de l'estimateur  $\hat{T}$  en utilisant le théorème 4.

$$\begin{aligned} \text{var}_M(\hat{T} - T) &= (N - n)\sigma^2 + \mathbf{1}'_r X_r G X'_r \mathbf{1}_r \sigma^2 \\ &= \sum_{i=1}^I (N_i - n_i)\sigma^2 + \sum_{i=1}^I (N_i - n_i)^2 / n_i \sigma^2 \\ &= \sum_{i=1}^I (N_i - n_i)\sigma^2 + \sum_{i=1}^I (N_i^2 - 2n_i N_i + n_i^2) / n_i \sigma^2 \\ &= \sum_{i=1}^I (N_i^2 / n_i) (1 - 2n_i / N_i + n_i^2 / N_i^2 + n_i / N_i - n_i^2 / N_i^2) \sigma^2 \\ &= \sum_{i=1}^I (N_i^2 / n_i) (1 - n_i / N_i) \sigma^2 \end{aligned}$$

Nous reconnaissons la variance dans le cas stratifié.

### 2.3.3 Estimation du total dans un modèle bifactoriel sans interaction

Dans ce modèle, il y a deux facteurs qui sont croisés. Les valeurs des  $Y_i$  sont mesurées pour tous les niveaux du premier facteur croisés avec chaque niveau du second facteur.

Toutes les cases constituées par le croisement des niveaux des facteurs doivent contenir au moins une unité échantillonnée. Le modèle sans interaction entre les facteurs s'écrit :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad (2.13)$$

où  $\varepsilon_{ij} \sim (0, \sigma^2)$ ,  $k \in \{1, \dots, N_{ij}\}$ ,  $i \in \{1, \dots, I\}$ , et  $j \in \{1, \dots, J\}$

Le nombre d'unités échantillonnées dans la case  $(i, j)$  est  $n_{ij}$ . Notons  $n_{i.} = \sum_{j=1}^J n_{ij}$  et  $n_{.j} = \sum_{i=1}^I n_{ij}$ . La taille de l'échantillon est  $n = n_{i.} + n_{.j} = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$ .  $Y_{s_{i.}}$  est la somme des  $Y_{ijk}$  pour le niveau  $i$  du facteur associé à  $\alpha$ .  $Y_{s_{.j}}$  est la somme des  $Y_{ijk}$  pour le niveau  $j$  du facteur associé à  $\beta$ . Comme on s'y attend  $Y_{s..}$  est la somme des  $Y_{ijk}$  pour tout l'échantillon. Avec ces notations l'équation normale s'écrit :

$$X'_s X_s \beta^0 = \begin{pmatrix} n & n_{1.} & n_{2.} & \dots & n_{I.} & n_{.1} & n_{.2} & \dots & n_{.J} \\ n_{1.} & n_{11} & 0 & 0 & 0 & n_{11} & n_{12} & \dots & n_{1J} \\ n_{2.} & 0 & n_{21} & 0 & 0 & n_{21} & n_{22} & \dots & n_{2J} \\ \vdots & 0 & 0 & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\ n_{I.} & 0 & 0 & 0 & n_{I1} & n_{I2} & \dots & n_{IJ} \\ n_{.1} & n_{11} & n_{12} & \dots & n_{1I} & n_{.1} & 0 & 0 & 0 \\ n_{.2} & n_{21} & n_{22} & \dots & n_{2I} & 0 & n_{.2} & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & 0 & 0 & \ddots & 0 \\ n_{.J} & n_{J1} & n_{J2} & \dots & n_{JI} & 0 & 0 & 0 & n_{.J} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_J \end{pmatrix}$$

$$= \begin{pmatrix} Y_{s..} \\ Y_{s_{1.}} \\ Y_{s_{2.}} \\ \vdots \\ Y_{s_{I.}} \\ Y_{s_{.1}} \\ Y_{s_{.2}} \\ \vdots \\ Y_{s_{.J}} \end{pmatrix} = X'_s Y_s$$

Ici  $X'_s X$  est d'ordre  $I + J + 1$  mais de rang  $I + J - 1$ , il y a donc deux colonnes qui sont combinaisons linéaires des autres. Pour résoudre l'équation normale, il n'est pas facile de trouver une sous matrice de dimension  $(I+J-1) \times (I+J-1)$  inversible. Il est donc souvent très difficile d'utiliser cette approche pour trouver l'estimateur du total et sa variance comme dans un modèle à un facteur. Searle (1971, sec. 7.1.d) décrit comment résoudre l'équation normale par la méthode qu'il appelle "absorption process". Dans cette approche les expressions pour  $\mu$  et un des  $\beta$  sont éliminées, les  $\alpha$  sont exprimés en fonction des  $\beta$  et une résolution analytique donne des  $\beta$  en fonction des  $Y_{s_{i.}}$  et des

$Y_{s,j}$ . Searle souligne cependant que cette approche ne marche pas à tous les coups. Toutefois il est toujours possible de calculer un g-inverse en utilisant la décomposition en valeurs singulières (voir p. 417 de Valliant, Dorfman et Royall (2000)). Il est alors facile d'évaluer numériquement l'estimateur du total et sa variance donné dans le théorème 4.

### 2.3.4 Estimation du total dans un modèle bifactoriel avec interaction

Le modèle avec interaction est numériquement plus facile à obtenir que celui sans interaction même s'il possède plus de paramètres à estimer. Le modèle avec interaction s'écrit :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}, \quad (2.14)$$

où  $\varepsilon_{ij} \sim (0, \sigma^2)$ ,  $k \in \{1, \dots, N_{ij}\}$ ,  $i \in \{1, \dots, I\}$ , et  $j \in \{1, \dots, J\}$

Il est assez difficile d'écrire la matrice  $X'_s$  dans le cas général. Pour une meilleure compréhension, nous considérerons l'exemple donné par le tableau de contingence suivant :

| i        | j=1 | j=2 | $n_{i.}$ |
|----------|-----|-----|----------|
| 1        | 2   | 2   | 4        |
| 2        | 2   | 1   | 3        |
| 3        | 1   | 3   | 4        |
| $n_{.j}$ | 5   | 6   | 11       |

TAB. 2.1 – Taille des cases formées par les niveaux des deux facteurs

Avec ce tableau de fréquence, il est possible d'écrire l'équation normale sous forme

matricielle.

$$\begin{pmatrix} Y_{s..} \\ Y_{s1..} \\ Y_{s2..} \\ Y_{s3..} \\ Y_{s.1.} \\ Y_{s.2.} \\ Y_{s11.} \\ Y_{s12.} \\ Y_{s21.} \\ Y_{s22.} \\ Y_{s31.} \\ Y_{s32.} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{31} \\ \gamma_{32} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{311} \\ \varepsilon_{321} \\ \varepsilon_{322} \\ \varepsilon_{323} \end{pmatrix}$$

Dans cet exemple nous avons 12 paramètres à estimer ( $\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \gamma_{31}, \gamma_{32}$ ) avec seulement 11 observations ce qui n'est pas possible. Cet exemple sert juste à illustrer la construction de la matrice  $X'_s$ . Il faudrait plus d'observations que de paramètres pour être en mesure de faire les estimations. Et dans la littérature, il existe des relations simples entre la taille d'échantillon et le nombre de paramètres. Cela dans le but de trouver la taille d'échantillon minimale qui garantit une estimation fiable des paramètres du modèle.

L'équation normale pour cet exemple s'écrit :

$$X'_s X_s \beta^0 = \begin{pmatrix} n & n_1. & n_2. & n_3. & n_{.1} & n_{.2} & n_{11} & n_{12} & n_{21} & n_{22} & n_{31} & n_{32} \\ n_1. & n_1. & 0 & 0 & n_{11} & n_{12} & n_{11} & n_{12} & 0 & 0 & 0 & 0 \\ n_2. & 0 & n_2. & 0 & n_{21} & n_{22} & 0 & 0 & n_{21} & n_{22} & 0 & 0 \\ n_3. & 0 & 0 & n_3. & n_{31} & n_{32} & 0 & 0 & 0 & 0 & n_{31} & n_{32} \\ n_{.1} & n_{11} & n_{21} & n_{31} & n_{.1} & 0 & n_{11} & 0 & n_{21} & 0 & n_{31} & 0 \\ n_{.2} & n_{12} & n_{22} & n_{32} & 0 & n_{.2} & 0 & n_{12} & 0 & n_{22} & 0 & n_{32} \\ n_{11} & n_{11} & 0 & 0 & n_{11} & 0 & n_{11} & 0 & 0 & 0 & 0 & 0 \\ n_{12} & n_{12} & 0 & 0 & 0 & n_{12} & 0 & n_{12} & 0 & 0 & 0 & 0 \\ n_{21} & 0 & n_{21} & 0 & n_{21} & 0 & 0 & 0 & n_{21} & 0 & 0 & 0 \\ n_{22} & 0 & n_{22} & 0 & 0 & n_{22} & 0 & 0 & 0 & n_{22} & 0 & 0 \\ n_{31} & 0 & 0 & n_{31} & n_{31} & 0 & 0 & 0 & 0 & 0 & n_{31} & 0 \\ n_{32} & 0 & 0 & n_{32} & 0 & n_{32} & 0 & 0 & 0 & 0 & 0 & n_{32} \end{pmatrix}$$

$$\times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{31} \\ \gamma_{32} \end{pmatrix} = \begin{pmatrix} Y_{s..} \\ Y_{s1..} \\ Y_{s2..} \\ Y_{s3..} \\ Y_{s.1.} \\ Y_{s.2.} \\ Y_{s11.} \\ Y_{s12.} \\ Y_{s21.} \\ Y_{s22.} \\ Y_{s31.} \\ Y_{s32.} \end{pmatrix} = X'_s Y_s.$$

Dans le cas général, la matrice  $X'_s X_s$  est carrée de dimension  $(1+I+J+IJ) \times (1+I+J+IJ)$  mais son rang est inférieur à  $(1+I+J+IJ)$ .  $IJ$  équations sont indépendantes. Dans l'exemple, la sous matrice inférieure 6x6 est inversible. Le rang de  $X'_s X_s$  est  $IJ = 2 * 3 = 6$ . Un g-inverse de  $X'_s X_s$  est donc  $G = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \text{diag}(1/n_{ij}) \end{pmatrix}$  où  $\mathbf{0}$  est la matrice formée de 0 de dimension  $(1+I+J) \times (1+I+J)$  dans le cas général, ce qui donne que  $\mathbf{0}$  est de dimension 6x6 dans cet exemple.

Une solution particulière de l'équation normale est

$$\beta^0 = (\mathbf{0}' \quad \bar{Y}'_s)'$$

où  $\mathbf{0}$  est le vecteur nul de longueur  $1 + I + J$  et  $\bar{Y}_s$  est le vecteur de longueur  $IJ$  des moyennes  $\bar{Y}_{s_{ij}}$ .

Le meilleur prédicteur linéaire non biaisé du total est alors :

$$\begin{aligned} \hat{T} &= \mathbf{1}'_s Y_s + \mathbf{1}'_r X_r \beta^0 \\ \hat{T} &= \sum_{i=1}^I \sum_{j=1}^J n_{ij} \bar{Y}_{s_{ij}} + \sum_{i=1}^I \sum_{j=1}^J (N_{ij} - n_{ij}) \bar{Y}_{s_{ij}} \\ \hat{T} &= \sum_{i=1}^I \sum_{j=1}^J N_{ij} \bar{Y}_{s_{ij}}. \end{aligned}$$

Et la variance est :

$$\begin{aligned}
\text{var}_M(\hat{T} - T) &= (N - n)\sigma^2 + 1'_r X_r G X'_r 1_r \sigma^2 \\
&= \sum_{i=1}^I \sum_{j=1}^J (N_{ij} - n_{ij})\sigma^2 + \sum_{i=1}^I \sum_{j=1}^J (N_{ij} - n_{ij})^2 / n_{ij} \sigma^2 \\
&= \sum_{i=1}^I \sum_{j=1}^J (N_{ij} - n_{ij})\sigma^2 + \sum_{i=1}^I \sum_{j=1}^J (N_{ij}^2 - 2n_{ij}N_{ij} + n_{ij}^2) / n_{ij} \sigma^2 \\
&= \sum_{i=1}^I \sum_{j=1}^J (N_{ij}^2 / n_{ij}) (1 - 2n_{ij} / N_{ij} + n_{ij}^2 / N_{ij}^2 + n_{ij} / N_{ij} - n_{ij}^2 / N_{ij}^2) \sigma^2 \\
&= \sum_{i=1}^I \sum_{j=1}^J (N_{ij}^2 / n_{ij}) (1 - n_{ij} / N_{ij}) \sigma^2.
\end{aligned}$$

On reconnaît de nouveau les estimateurs de la moyenne et de la variance dans le cas d'un modèle avec deux variables de stratification. Nous constatons que pour faire les estimations, nous avons besoin de connaître la taille de la population dans chacune des cellules formées par les niveaux des deux facteurs.

## 2.4 Modèles avec variables continues

Jusqu'à présent nous avons considéré des modèles avec seulement des facteurs. Cependant des variables continues appelées covariables peuvent être ajoutées au modèle. L'introduction des covariables permet d'augmenter la précision du modèle et en analyse de la variance, c'est une façon de réduire l'effet nuisible de certaine variable incontrôlable. Le fait d'introduire des covariables n'a aucune incidence sur l'estimation des paramètres déjà vue dans les sections précédentes. On pourra donc utiliser la technique de l'inverse généralisé pour trouver les paramètres du modèle et les quantités à estimer comme le total ou la moyenne.

### 2.4.1 Modèle général avec covariance

Soit  $X$  une matrice de dimension  $N \times p$  de variables auxiliaires qualitatives et  $Z$  une matrice de dimension  $N \times q$  de variables quantitatives ou covariables. On peut écrire le modèle de la façon suivante :

$$Y = X\beta + Z\gamma + \varepsilon, \quad (2.15)$$

où  $\beta$  est un vecteur de longueur  $p$ ,  $\gamma$  est un vecteur de longueur  $q$  et  $\varepsilon$  est un vecteur de longueur  $N$  avec moyenne  $\mathbf{0}$  et variance  $\sigma^2 I$ . On fait l'hypothèse que  $X$  n'est pas nécessairement de plein rang mais que  $Z$  est de plein rang. Mais aussi que les colonnes de  $Z$  ne peuvent pas être exprimées comme des combinaisons linéaires de celles de  $X$  et que la matrice  $X_s$  des unités échantillonnées n'est pas de plein rang. Donc  $X'_s X_s$  n'est pas inversible par contre  $Z'_s Z_s$  a un inverse. L'équation normale s'écrit :

$$\begin{pmatrix} X'_s X_s & X'_s Z_s \\ Z'_s X_s & Z'_s Z_s \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} X'_s Y_s \\ Z'_s Y_s \end{pmatrix} \quad (2.16)$$

Une solution de l'équation normale est :

$$\begin{pmatrix} \beta \\ \gamma \end{pmatrix} = G^* \begin{pmatrix} X'_s Y_s \\ Z'_s Y_s \end{pmatrix} \quad (2.17)$$

où  $G^*$  est un g-inverse de  $\begin{pmatrix} X'_s X_s & X'_s Z_s \\ Z'_s X_s & Z'_s Z_s \end{pmatrix}$ . Mais il est aussi possible de résoudre l'équation normale (2.16) comme deux équations à deux inconnues  $\beta$  et  $\gamma$ . La résolution de la première équation donne :

$$\beta^0 = G_x (X'_s Y_s - X'_s Z_s \gamma^0) = \beta^* - G_x X'_s Z_s \gamma^0, \quad (2.18)$$

où  $G_x$  est un g-inverse de  $X'_s X_s$  et  $\beta^* = G_x X'_s Y_s$ .  $\beta^*$  est une solution du modèle sans covariables  $Z$ . En remplaçant la solution dans l'équation 2, on a :  $Z'_s P_s Z_s \gamma^0 = Z'_s P_s Y_s$  où  $P_s = I_n - X_s G_x X'_s$ . Donc

$$\gamma^0 = G_1 Z'_s P_s Y_s, \quad (2.19)$$

où  $G_1$  est un g-inverse de  $Z'_s P_s Z_s$ . La solution  $\gamma^0$  est unique, on pouvait s'y attendre vu que la matrice  $Z$  est de plein rang c'est à dire  $Z'_s Z_s$  possède un inverse unique. Pour le montrer, donnons des éléments de réponse. Dans la solution  $\gamma^0$ , on retrouve  $G_x$  qui n'est pas unique mais  $X_s G_x X'_s$  est invariant au choix de  $G_x$  g-inverse de  $X'_s X_s$ . Donc le fait que  $X'_s X_s$  ne soit pas inversible, va pas engendrer une non unicité de la solution  $\gamma^0$ . Le facteur  $P_s$  est symétrique et idempotent c'est à dire que  $P_s^2 = P_s$  et  $(Z'_s Z_s)^{-1}$  existe. Ces propriétés permettent de montrer que  $Z'_s P_s Z_s$  est non singulière (Searle, 1971, sec. 8.2). La solution unique de l'estimateur de  $\gamma$  est :

$$\hat{\gamma} = (Z'_s P_s Z_s)^{-1} Z'_s P_s Y_s. \quad (2.20)$$

L'estimateur du total est :

$$\hat{T} = 1'_s Y_s + 1'_r X_r \beta^0 + 1'_r Z_r \hat{\gamma}.$$

La variance de l'estimateur est donné par le lemme suivant :

**Lemme 5.** La variance de  $\hat{T} = 1'_s Y_s + 1'_r X_r \beta^0 + 1'_r Z_r \hat{\gamma}$  sous le modèle linéaire général (1.2) est :

$$\text{var}_M(\hat{T} - T) = 1'_r V_{rr} 1_r + 1'_r (X_r Z_r) \text{var}_M \begin{pmatrix} \beta^0 \\ \hat{\gamma} \end{pmatrix} \begin{pmatrix} X'_r \\ Z'_r \end{pmatrix} 1_r, \quad (2.21)$$

$$\text{où } \text{var}_M \begin{pmatrix} \beta^0 \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} G_x + G_x X'_s Z_s A_z^{-1} Z'_s X_s G'_x & -G_x X'_s Z_s A_z^{-1} \\ -A_z^{-1} Z'_s X_s G'_x & A_z^{-1} \end{pmatrix} \sigma^2$$

et  $A_z = Z'_s P_s Z_s$ .

### Preuve du lemme (5) :

D'après l'expression de  $\beta^0$  :  $\text{Var}_M(\beta^0) = G_x X'_s Q Q' X_s G_x \sigma^2$  où  $Q = I_n - Z_s A^{-1} Z'_s P_s$ . Par le lemme (3)(ii), il vient que  $X'_s P_s = 0$ .

Comme  $GA_s G' = G$  donc  $G_x X'_s Q Q' X_s G_x = G_x + G_x X'_s Z_s A_z^{-1} Z'_s X_s G'_x$  d'où le résultat.  $\text{cov}_M(\beta^0, \hat{\gamma}) = G_x X'_s Q P_s Z_s A^{-1} \sigma^2$  si on développe  $Q P_s$  et on utilise  $GA_s G' = G$  on obtient que  $\text{cov}_M(\beta^0, \hat{\gamma}) = -G_x X'_s Z_s A_s^{-1}$ .

## 2.4.2 Modèle unifactoriel avec une covariable

Ce cas particulier va permettre d'appliquer les formules générales précédemment trouvées. Nous considérons le modèle avec un facteur et une covariable. Supposons que le facteur possède  $I$  niveaux et notons  $Z_{ij}$  la variable quantitative pour le niveau  $j$  de l'unité  $i$ . Le reste de la notation est similaire au modèle avec un facteur (2.10). Le modèle s'écrit alors :

$$Y_{ij} = \mu + \beta_j + \gamma z_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim (0, \sigma^2) \quad (2.22)$$

Les erreurs sont supposées indépendantes comme dans ce qui précède.  $X$  a la même forme que dans le modèle unifactoriel.  $\beta^* = (0, \bar{Y}_{s_1}, \dots, \bar{Y}_{s_I})$  est la solution pour un modèle unifactoriel sans covariable. En utilisant la définition de  $X_s$  et  $G_x$ , on a :

$$P_s = I - \text{blocdiag}(n_i^{-1} \mathbf{1}_{n_i} \mathbf{1}'_{n_i})$$

où  $\mathbf{1}_{n_i}$  est un vecteur de longueur  $n_i$  avec que des 1. Le vecteur des variables quantitatives échantillonnées est :  $Z_s = [z_{11}, \dots, z_{1n_1}, \dots, z_{I1}, \dots, z_{In_I}]$ . L'estimateur de  $\gamma$  est :

$$\hat{\gamma} = \frac{\sum_{i=1}^I \sum_{j \in s_i} (z_{ij} - \bar{z}_{s_i}) Y_{ij}}{\sum_{i=1}^I \sum_{j \in s_i} (z_{ij} - \bar{z}_{s_i})^2} \quad (2.23)$$

De même  $G_x X'_s Z'_s = (0, \bar{z}_{s_1}, \dots, \bar{z}_{s_I})$  où  $\bar{z}_{s_i} = \sum_{j \in s_i} z_{ij}/n_i$ . En remplaçant ce résultat dans l'expression de  $\beta^0$  (2.18), on trouve :

$$\beta^0 = \beta^* - G_x X'_s Z'_s \gamma^0 = \begin{pmatrix} \bar{Y}_{s_1} - \hat{\gamma} \bar{z}_{s_1} \\ \vdots \\ \bar{Y}_{s_I} - \hat{\gamma} \bar{z}_{s_I} \end{pmatrix} \quad (2.24)$$

Donc on obtient comme estimateur du total après un peu d'arrangement :

$$\hat{T} = \sum_{i=1}^I N_i [\bar{Y}_{s_i} + \hat{\gamma}(\bar{z}_i - \bar{z})].$$

Cet estimateur peut aussi s'écrire sous la forme

$$\hat{T} = \sum_{i=1}^I \sum_{j \in s_i} N_i [1/n_i + A^{-1}(\bar{z}_i - \bar{z}_{s_i})(z_{ij} - \bar{z}_{s_i})].$$

où  $A_z = \sum_{i=1}^I \sum_{j \in s_i} (z_{ij} - \bar{z}_{s_i})$ . Cette formule est pratique pour calculer la variance.

$$\hat{T} - T = \sum_{i=1}^I \sum_{j \in s_i} (N_i d_{ij} - 1) Y_{ij} - \sum_{i=1}^I \sum_{j \in s_i} Y_{ij}$$

où  $d_{ij} = 1/n_i + A_z^{-1}(\bar{z}_i - \bar{z}_{s_i})(z_{ij} - \bar{z}_{s_i})$  et  $r_i$  représente les unités du niveau  $i$  non échantillonnées. La variance de l'estimateur est alors  $var(\hat{T} - T) = \sum_{i=1}^I \sum_{j \in s_i} (N_i d_{ij} - 1)^2 \sigma^2 - \sum_{i=1}^I \sum_{j \in s_i} \sigma^2$  avec  $\sum_{j \in s_i} d_{ij}^2 = 1/n_i + A_z^{-1}(\bar{z}_i - \bar{z}_{s_i})^2$  et  $\sum_{j \in s_i} d_{ij} = 1$ . Donc

$$var(\hat{T} - T) = \sigma^2 \sum_{i=1}^I \frac{N_i^2}{n_i} (1 - f_i) \left[ 1 + \frac{n_i}{(1 - f_i) A_z} (\bar{z}_i - \bar{z}_{s_i})^2 \right].$$

Ce résultat est très similaire à celui de la régression linéaire simple du chapitre précédent (exemple 2). Notons que l'échantillon équilibré d'ordre 1 pour tous les niveaux  $i$  c'est à dire  $\bar{z}_i = \bar{z}_{s_i} \quad \forall i \in \{1, \dots, I\}$  est optimal.

# Chapitre 3

## Simulation

Cette partie est consacrée à une petite étude de simulation. On s'intéresse à une population caractérisée par deux variables qualitatives. Pour tout croisement des niveaux des deux facteurs, nous avons une et une seule réalisation de  $Y$ . Cette situation n'est pas beaucoup traitée dans la littérature. Le but est de comparer les estimations selon le plan de sondage avec ceux par le modèle. Pour cela, nous simulerons une population de 900 individus selon trois scénarios différents (normale, autorégressif d'ordre 1, cyclique). Pour chacune de ces trois populations, nous effectuons trois types d'échantillonnage du plus standard, l'aléatoire simple, à un non standard, en passant par l'échantillonnage stratifié. Au total, neuf estimations par rapport au plan de sondage vont être comparées à leur équivalent en utilisant le modèle. La racine carrée de l'erreur quadratique moyenne pris relativement à la moyenne est notre critère de comparaison. Pour des estimateurs non biaisés, la racine carrée de l'erreur quadratique moyenne relativement à la moyenne est une estimation du coefficient de variation.

### 3.1 Population

On choisit  $I = J = 30$  niveaux pour les variables auxiliaires  $X_1$  et  $X_2$ . Il est possible de voir les données sous une structure de tableau. Nous considérerons que les niveaux de  $X_1$  sont en ligne et ceux de  $X_2$  en colonne. En d'autres termes, chaque ligne représente un niveau de  $X_1$  et chaque colonne est un niveau de  $X_2$ . D'où les expressions effet de ligne ou effet de colonne. Pour toute la population il y a donc 30 lignes et 30 colonnes. Comme chaque cellule  $(i, j)$  ne contient qu'une unité  $Y_{ij}$ , la population compte  $30 * 30 = 900$  individus.

|          |           | $X_2$    |           |          |            |           |
|----------|-----------|----------|-----------|----------|------------|-----------|
|          |           | 1        | ...       | j        | ...        | 30        |
| $X_1$    | 1         | $Y_{11}$ | ...       | $Y_{1j}$ | ...        | $Y_{130}$ |
|          | $\vdots$  | $\vdots$ | $\vdots$  | $\vdots$ | $\vdots$   | $\vdots$  |
|          | i         | $Y_{i1}$ | ...       | $Y_{ij}$ | ...        | $Y_{i30}$ |
| $\vdots$ | $\vdots$  | $\vdots$ | $\vdots$  | $\vdots$ | $\vdots$   | $\vdots$  |
| 30       | $Y_{301}$ | $\vdots$ | $Y_{30j}$ | ...      | $Y_{3030}$ |           |

TAB. 3.1 – forme en grille de la population à simuler

Le modèle considéré est le modèle linéaire général sans interaction :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{où} \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \quad \text{et} \quad \sum \alpha_i = \beta_j = 0. \quad (3.1)$$

Pour générer la population, on fixe  $\mu$  et on génère les effets  $\alpha_i$ ,  $\beta_j$  et les erreurs  $\varepsilon_{ij}$ . On sait que les erreurs suivent une loi normale avec moyenne 0 et variance  $\sigma_\varepsilon^2$ . Il reste donc à simuler les effets  $\alpha_i$  et  $\beta_j$  de respectivement  $X_1$  et  $X_2$ . Pour ce faire, nous choisissons 3 fonctions. Chacune de ces fonctions permettra d'obtenir une population différente. Ainsi il y aura 3 populations distinctes sur lesquelles la même étude va porter.

### 3.1.1 Première fonction : Loi Normale

Dans ce cas les effets sont considérés comme suivant une distribution normale. Nous avons :

$$\alpha_i \sim N(0, \sigma_\alpha^2) \quad \text{et} \quad \beta_j \sim N(0, \sigma_\beta^2). \quad (3.2)$$

Les variances  $\sigma_\alpha^2$  et  $\sigma_\beta^2$  sont plus élevés que la variabilité des erreurs  $\sigma_\varepsilon^2$ . De plus on choisira  $\sigma_\beta^2$  plus élevé que  $\sigma_\alpha^2$ , ce qui sera une justification partielle de la stratification selon  $X_2$  plus loin dans la simulation. Tous les effets sont en moyenne nuls. Selon cette fonction, les effets de ligne ou de colonne sont en moyenne équivalents.

### 3.1.2 Deuxième fonction : Processus Autorégressif d'ordre 1

Ici les effets sont générés selon une loi autorégressive d'ordre 1. Nous avons :

$$\begin{aligned} \alpha_i &= a * \alpha_{i-1} + N(0, \sigma_a^2), \\ \beta_j &= b * \beta_{j-1} + N(0, \sigma_b^2). \end{aligned} \quad (3.3)$$

où  $|a|$  et  $|b|$  sont inférieures à 1,  $\alpha_1 \sim N(0, \sigma_a^2/(1-a^2))$  et  $\beta_1 \sim N(0, \sigma_b^2/(1-b^2))$ . Comme  $|a|$  et  $|b|$  sont inférieures à 1 donc les 2 processus autorégressifs convergent. On peut déduire les variances des effets :

$$\text{var}(\alpha_i) = \sigma_a^2/(1-a^2) \quad \text{et} \quad \text{var}(\beta_j) = \sigma_b^2/(1-b^2). \quad (3.4)$$

Cette fonction traduit le fait que l'effet d'une ligne est une fraction de l'effet de la ligne précédente. Le même raisonnement est valable pour les colonnes.

### 3.1.3 Troisième fonction : Processus Cyclique

Ici les effets sont générés selon la fonction sinus. Cette fonction est cyclique avec une période de  $2\pi$ . On s'arrange pour que les effets de  $X_1$  s'étale sur un cycle, cela s'applique à  $X_2$  également. D'où

$$\begin{aligned} \alpha_i &= \sin(2\pi i/30) + N(0, \sigma_\alpha^2), \\ \beta_j &= \sin(2\pi j/30) + N(0, \sigma_\beta^2). \end{aligned} \quad (3.5)$$

Ici 30 est le nombre de niveaux pour les variables qualitatives  $X_1$  et  $X_2$ . Il vient donc que les variances des effets sont :

$$\text{var}(\alpha_i) = E[\sum (\alpha_i - \bar{\alpha})^2/29] \quad \text{et} \quad \text{var}(\beta_j) = E[\sum (\beta_j - \bar{\beta})^2/29]. \quad (3.6)$$

En moyenne la première moitié des lignes ont des effets positifs et l'autre moitié des effets négatifs. Si on regroupe les niveaux de  $X_1$  en 4 groupes à partir du début. Dans chaque groupe, plus les niveaux sont éloignés plus leur effet est différent. Le raisonnement tient aussi pour les colonnes.

Pour rendre les paramètres estimables, nous centrons tous les effets. Ce qui donne

$$\alpha_i^* = \alpha_i - \bar{\alpha} \quad \text{et} \quad \beta_j^* = \beta_j - \bar{\beta}.$$

où  $\alpha_i$  et  $\beta_j$  sont les effets respectifs du niveau  $i$  de  $X_1$  et du niveau  $j$  de  $X_2$  obtenus à partir des processus normal, autorégressif et cyclique.  $\alpha_i^*$  et  $\beta_j^*$  sont les valeurs centrées des effets, utilisées dans l'inférence pour l'estimation des paramètres du modèle. Dans la suite pour des raisons d'allègement de la notation,  $\alpha_i$  et  $\beta_j$  désigneront ces valeurs centrées.

Les paramètres de la population comme la moyenne globale  $\mu$ , les variances des effets  $\sigma_\alpha^2$  et  $\sigma_\beta^2$  sont choisis de sorte que le coefficient de variation ( $CV$ ) soit élevé. Dans notre cas, nous voulons un  $CV$  dans la population supérieur à 10%. Le coefficient de

variation est une mesure relative de la variation, c'est le rapport entre l'écart type et la moyenne c'est à dire  $CV = \sigma/\mu$ . Cette mesure est donc sans unité. Les données avec une variabilité grande sont plus intéressantes à échantillonner comparativement à des données qui bougent relativement peu autour de leur moyenne.

Le programme R qui a servi à faire la simulation de la population se trouve dans l'annexe C. Le nom de la fonction est `population()`.

## 3.2 Échantillonnage

Pour chacune des 3 populations simulées, nous effectuons 3 types d'échantillonnages. Il s'agit de l'échantillonnage aléatoire simple (*AS*), de l'échantillonnage stratifié (*ST*) et d'un type d'échantillonnage non standard (*NS*). Ce dernier consiste à tirer de façon aléatoire un couple de niveaux dont le premier provient de  $X_1$  et le second de  $X_2$ . Tous les échantillons tirés sont de tailles  $n = 120$ . Ce qui donne une fraction de sondage de  $f = n/N = 120/900 \simeq 13,33\%$ . Le programme R est dans l'annexe C sous le nom `tirage()`.

### 3.2.1 Échantillonnage Aléatoire Simple

Dans cette forme d'échantillonnage, toutes les unités ont la même probabilité d'être tirées. Nous ne tenons donc pas compte de la structure ligne  $\times$  colonne (en grille) des données. Une façon de faire est de numéroter les individus de 1 à  $30 * 30 = 900$  et de choisir  $n = 120$  au hasard parmi ces 900 sans remise. Il est donc impossible de choisir plus d'une fois un individu. Il s'agit ensuite de récupérer les niveaux de  $X_1$  et de  $X_2$  des 120 unités échantillonnées et d'estimer les effets de ces niveaux. Malheureusement de par la méthode de tirage et de la faible fraction de sondage, il arrive assez souvent qu'au moins un effet de ligne ou de colonne ne soit pas estimable car la ligne ou la colonne correspondante n'a pas été tirée. Dire qu'une ligne ou une colonne n'est pas tirée signifie qu'aucun des individus tirés ne se situe sur la ligne ou la colonne en question. Nous ne pouvons pas alors estimer l'effet de la ligne ou de la colonne manquante. Dans une telle situation, nous ignorons les lignes ou les colonnes dont l'effet n'est pas estimable. Le taux de non estimé est la proportion de fois où au moins une ligne ou une colonne n'a été tirée.

### 3.2.2 Échantillonnage Stratifié

Les strates sont les colonnes, donc le facteur  $X_2$  est la variable de stratification. Cela du fait que la variabilité des effets colonnes est plus élevée que celle des effets lignes. Comme il y a 30 colonnes, il s'agit de choisir au hasard 4 individus sur chaque colonne. De ce fait la taille d'échantillon est identique dans toutes les strates ( $n_h = 4$ ). Ici tous les effets colonnes sont estimés car aucune colonne n'est omise. Par contre certaines lignes peuvent ne pas être tirées. Contrairement au cas aléatoire simple, c'est toujours des lignes qui ne sont pas estimables cela peut être doublement problématique. En effet le biais dû à la non estimabilité se produit à chaque fois du même côté (ligne) et de plus pour un taux de "non estimé" équivalent entre aléatoire simple et stratifié, on peut penser qu'en moyenne deux fois plus souvent les lignes ne sont pas estimées dans le cas stratifié comparativement au cas aléatoire simple. Imaginons que les effets de ligne soient très distincts, l'estimation peut être beaucoup affectée si des effets de ligne sont très souvent non estimés c'est à dire absents.

Cette situation serait préoccupante si les effets de ligne ne suivaient pas la même loi que les effets de colonne. Ce qui n'est pas le cas ici. La précision des estimations dépend surtout du taux global d'effets non estimés.

### 3.2.3 Échantillonnage Non Standard

De par la méthode d'échantillonnage, tous les effets de ligne et de colonne sont estimés car toutes les lignes et colonnes sont tirées. L'algorithme consiste à créer 2 vecteurs de longueur  $n = 120$  dont les niveaux 1 à 30 sont répliqués 4 fois chacun. Les 2 vecteurs représentent les niveaux des deux facteurs  $X_1$  et  $X_2$  pour les individus échantillonnés. On fait une permutation aléatoire de chaque vecteur. Les couples des 2 vecteurs permutés définissent les unités tirées. Nous constatons comme mentionné précédemment que tous les effets de ligne et de colonne vont être estimables mais une unité peut être tirée plus d'une fois. Dans ce cas le poids des individus n'est pas identique. Nous pourrions laisser l'échantillon tel quel ou alors décider d'éliminer les répétitions et de ce fait réduire la taille de l'échantillon. Dans tous les cas les effets sont estimés en totalité, on peut donc s'attendre à ce que l'estimation de la moyenne ou du total soit meilleure dans le cas non standard en comparaison avec l'échantillonnage aléatoire simple ou stratifié.

## 3.3 Inférence et résultats

### 3.3.1 Inférence

Pour atteindre des résultats asymptotiques, 500 échantillons pour chacun des 9 cas de figure (3 populations  $\times$  3 types d'échantillonnage) ont été tirés. Ce processus est répété 10 fois. Les résultats finaux sont une moyenne des 10 estimations issues des groupes de 500 échantillons. Au total, nous avons généré 5000 échantillons pour faire l'inférence. Pour chaque échantillon, nous procédons à l'estimation de la moyenne selon les deux approches.

La première approche est celle par le plan d'échantillonnage. Les formules de calcul de la moyenne selon les différentes méthodes de sélection (aléatoire, stratifié...) sont bien connues. Comme les individus ont tous le même poids,  $\hat{y} = (1/N) \sum_{i \in s} w_i y_i = \sum_{i \in s} y_i / n$  où  $w_i = N/n$  est le poids de l'individu  $i$ .

La seconde approche est celle par le modèle. Le modèle linéaire général (3.1) :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{où} \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2).$$

est celui qui est considéré ici. L'échantillon tiré permet d'estimer les effets de ligne et de colonne. Grâce à ces estimateurs, les  $Y_{ij}$  des individus  $(i, j)$  non tirés vont être prédits par le modèle (3.1) et notés  $\hat{y}_{ij}$ . L'estimateur de la moyenne est la moyenne de toutes les valeurs de  $Y$ , observés ( $Y_{ij}$ ) et prédites ( $\hat{Y}_{ij}$ ). Il arrive que des effets ligne ou colonne ne soient pas estimés du fait qu'aucun individu de la ligne ou de la colonne correspondante n'a été tiré. Dans une telle situation, nous choisissons d'éliminer cette ligne ou cette colonne de nos données à prédire. Ce qui va bien sûr affecter la précision de nos estimateurs. Rappelons que dans le cas de l'échantillonnage non standard, tous les effets de ligne ou de colonne sont obligatoirement estimés. Donc ce problème ne se pose pas par contre il arrive que des unités soient tirées plusieurs fois. Nous ne compterons qu'une fois les unités qui apparaissent plusieurs fois. La fraction de sondage peut donc être inférieure à  $120/900 = 0.1333$  soit 13,33%. Cette baisse signifie-t-elle une variabilité plus grande de notre estimateur par rapport aux deux autres? Le type d'échantillonnage non standard, dont il est question, est plus complexe. Il faudrait un travail de recherche plus approfondi sur la variance de son estimateur par rapport au plan de sondage et au modèle pour répondre à la question.

La racine carrée de l'erreur quadratique moyenne divisée par la moyenne dans la population va permettre de comparer les estimateurs par rapport au plan d'échantillonnage à ceux dérivés du modèle. Dans le cas des estimateurs non biaisés, cette quantité est l'estimateur du coefficient de variation de la moyenne. Évidemment, plus elle est petite plus l'estimateur est précis.

### 3.3.2 Résultats

Le modèle linéaire général (3.1) :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{où} \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2).$$

a été simulé avec les paramètres suivants :

- la variance des erreurs est  $\sigma_\varepsilon^2 = 0.5$
- la loi suivie par les effets sont :  
Dans le cas normal, nous avons :

$$\begin{aligned} \alpha_i &\sim N(0, 1.00) &\Rightarrow & \text{var}(\alpha_i) = 1.00 \\ \beta_i &\sim N(0, 1.25) &\Rightarrow & \text{var}(\beta_i) = 1.25 \end{aligned}$$

Dans le cas autorégressif d'ordre 1, nous avons :

$$\begin{aligned} \alpha_i &= 0.9 * \alpha_{i-1} + N(0, 1.00) &\Rightarrow & \text{var}(\alpha_i) = 1.00/(1 - 0.9^2) \approx 5.26 \\ \beta_i &= 0.9 * \beta_{i-1} + N(0, 1.25) &\Rightarrow & \text{var}(\beta_i) = 1.25/(1 - 0.9^2) \approx 6.58 \end{aligned}$$

Dans le cas cyclique, nous avons :

$$\begin{aligned} \alpha_i &= \sin(2\pi i/30) + N(0, 1.00) &\Rightarrow & \text{var}(\alpha_i) \approx 1.29 \\ \beta_i &= \sin(2\pi i/30) + N(0, 1.25) &\Rightarrow & \text{var}(\beta_i) \approx 1.63 \end{aligned}$$

- la moyenne globale  $\mu = 5$

Avec ces paramètres nous simulons une population de 900 unités avec la fonction *population()* voir Annexe C.

Avant de faire l'échantillonnage, nous vérifions que la population possède bien les paramètres voulus. Pour cela, nous utilisons la fonction *lm()* du logiciel R pour ajuster un modèle linéaire à deux facteurs avec les données de la population obtenue  $Y_{ij}$  où  $i, j \in 1, \dots, 30$ . Le MSE (mean square error) de la table ANOVA est une estimation non biaisée de la variance des erreurs  $\sigma_\varepsilon^2$ . Le tableau 3.2 donne le MSE obtenu dans chacune des trois populations. La variance théorique des erreurs est  $\sigma_\varepsilon^2 = 0.5$ . Les trois MSE obtenus sont très proches de 0.50. Nous remarquons que le population générée à partir de AR(1) est légèrement plus variable avec un MSE de 0.57 au lieu de 0.50. On peut également estimer  $\sum \alpha_i^2/(a - 1)$  et  $\sum \beta_j^2/(b - 1)$  car

$$\begin{aligned} E(MS_{\text{ligne}}) &= \sigma_\varepsilon^2 + (b/(a - 1)) \sum_{i=1}^{30} \alpha_i^2 \\ E(MS_{\text{colonne}}) &= \sigma_\varepsilon^2 + (a/(b - 1)) \sum_{j=1}^{30} \beta_j^2 \end{aligned}$$

Donc nous pouvons estimer la variabilité inter effets par :

$$\begin{aligned}\sum \alpha_i^2/(a-1) &= (1/b)[E(MSligne) - \sigma_\varepsilon^2] \\ \sum \beta_i^2/(b-1) &= (1/a)[E(MScolonne) - \sigma_\varepsilon^2]\end{aligned}$$

car  $\mu_\alpha = \mu_\beta = 0$ .

|                         | Normale | AR(1)  | Cyclique |
|-------------------------|---------|--------|----------|
| MSE                     | 0.4803  | 0.5652 | 0.4917   |
| $\sum \alpha_i^2/(a-1)$ | 1.0082  | 1.8836 | 1.4793   |
| $\sum \beta_i^2/(b-1)$  | 1.3072  | 3.0619 | 1.6391   |

TAB. 3.2 – Vérification des paramètres de nos populations

La variabilité estimée des effets coïncide avec les valeurs théoriques sauf dans le cas de AR(1). Cela est normal car les moyennes échantillonales  $\bar{\alpha}$  et  $\bar{\beta}$  sont non nulles. Donc  $var(\alpha_i) \neq \sum \alpha_i^2/(a-1)$  et  $var(\beta_i) \neq \sum \beta_i^2/(a-1)$ . La vérification est concluante en ce sens que les paramètres sont très proches de ceux anticipés.

On rappelle que pour faire les calculs de précision 5000 échantillons ont été réalisés. Il arrive dans le cas de l'échantillonnage aléatoire simple et stratifié que tous les effets ne soient pas estimés. Le tableau 3.3 qui suit donne le pourcentage d'échantillons dans lesquels tous les effets de ligne et de colonne sont estimés en fonction des trois types d'échantillonnage aléatoire simple (AS), stratifié (ST) et non standard (NS).

C'est dans le cas aléatoire simple qu'il y a plus d'échantillons avec des effets non es-

|             | AS     | ST     | NS    |
|-------------|--------|--------|-------|
| pourcentage | 44.6 % | 65.8 % | 100 % |

TAB. 3.3 – Pourcentage des échantillons dans lesquels tous les effets sont estimés.

timés. Cependant tous les effets non estimés dans le cas stratifié sont des effets de ligne car la stratification se fait sur les colonnes. Dans le cas non standard tous les effets sont estimés car toutes les lignes et toutes les colonnes sont tirées. Pour traiter ce problème de non estimabilité des effets, nous avons choisi de tout simplement éliminer les lignes ou les colonnes non tirées. Évidemment il serait possible de considérer des méthodes d'imputation pour estimer ces effets. On peut par exemple prendre la moyenne des effets immédiatement voisins.

Qu'en est-il du biais des estimateurs ? Le tableau 3.4 donne la moyenne des 500 biais divisée par la moyenne de la variable réponse pour toute la population. Les biais relatifs moyens sont très faibles de l'ordre du millième et moins. Nous pouvons sans ambiguïté conclure que les estimateurs de la moyenne, dont il est question ici, aussi bien par rapport au plan que par le modèle sont non biaisés.

|          | AS      |         | ST      |         | NS      |         |
|----------|---------|---------|---------|---------|---------|---------|
|          | Plan    | Modèle  | Plan    | Modèle  | Plan    | Modèle  |
| Normale  | 0.0121  | 0.0452  | -0.0279 | 0.0299  | -0.0056 | -0.0055 |
| Ar(1)    | -0.0772 | -0.0105 | -0.0271 | -0.0005 | -0.0023 | -0.0027 |
| Cyclique | 0.0485  | -0.0450 | 0.0789  | -0.0129 | -0.0110 | 0.0111  |

TAB. 3.4 – Moyennes des 10 Biais relatifs des estimateurs en pourcentage.

Du fait que les estimateurs sont non biaisés, la racine carrée des erreurs quadratiques moyennes divisée par la moyenne de la variable réponse dans la population est un bon estimateur du coefficient de variation. Le tableau 3.5 résume les résultats obtenus en pourcentage. On voit que les estimateurs sont précis les valeurs allant d'environ 1.23 % à 3.89 %. L'estimation par le modèle est meilleure que celle par le plan. Sauf dans le cas non standard où les différences sont non significatives. Dans ce dernier cas, l'estimation par le modèle est équivalente à celle par le plan. Si on compare les types d'échantillonnage, il vient que l'échantillonnage non standard est le plus précis des trois suivi par l'échantillonnage stratifié.

|          | AS   |        | ST   |        | NS   |        |
|----------|------|--------|------|--------|------|--------|
|          | Plan | Modèle | Plan | Modèle | Plan | Modèle |
| Normale  | 2.82 | 1.84   | 2.10 | 1.56   | 1.26 | 1.26   |
| Ar(1)    | 3.89 | 1.96   | 2.65 | 1.57   | 1.23 | 1.24   |
| Cyclique | 3.19 | 1.84   | 2.39 | 1.66   | 1.25 | 1.25   |

TAB. 3.5 – Moyenne des 10 estimations de CV en pourcentage.

### 3.3.3 Quand le modèle est "faux"

Ici nous essayons de voir ce qu'il advient des résultats lorsque le modèle linéaire général (3.1) ne tient pas. Dans cette section le vrai modèle est le multiplicatif :

$$\log(Y_{ij}) = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{où} \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2). \quad (3.7)$$

Pour ce qui est de la simulation de la population, les effets de ligne et de colonne ainsi que les erreurs sont générées de la même façon que précédemment avec les mêmes valeurs de paramètres. Par contre  $Y_{ij} = \exp(\mu + \alpha_i + \beta_j + \varepsilon_{ij})$  où  $i, j \in 1, \dots, 30$ . La variabilité des  $Y_{ij}$  va être plus grande que dans le cas précédent. Cela peut causer beaucoup de problèmes pour l'obtention de résultats asymptotiques.

Pourquoi dit-on que le modèle est "faux" ? Le principe est le même que dans les chapitres 1 et 2. Pour faire la prédiction de la variable  $Y$ , nous supposons le modèle linéaire général (3.1)  $\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$ . Alors que les données ne suivent pas ce modèle mais bien le modèle multiplicatif (3.7).

En pratique le vrai modèle n'est pas connu ou est trop complexe. Il est donc intéressant de savoir à quel point nos résultats vont être affectés par l'utilisation d'un modèle "faux". Le tableau 3.6 donne les biais relatifs moyens des estimateurs en pourcentage.

|          | AS      |         | ST      |          | NS      |         |
|----------|---------|---------|---------|----------|---------|---------|
|          | Plan    | Modèle  | Plan    | Modèle   | Plan    | Modèle  |
| Normale  | -0.0804 | -0.2737 | -0.2453 | 0.2874   | -0.4577 | -0.4587 |
| Ar(1)    | -0.4428 | -0.3849 | -0.0829 | -0.3140  | -0.4921 | -0.4888 |
| Cyclique | 0.0632  | -0.3352 | 0.0424  | 517.3408 | -0.5296 | -0.5299 |

TAB. 3.6 – Moyenne des 10 biais des estimateurs quand le modèle est "faux" en pourcentage.

Le modèle semble éprouver beaucoup de difficultés dans le cas cyclique avec un échantillonnage stratifié. En effet, le dernier cas montre un biais relatif moyen de 517.34 %. Il faut prendre ce résultat avec beaucoup de précaution. Quand nous regardons l'estimation des 10 biais séparément, nous constatons que le biais de l'échantillon 6 est au moins 10 fois plus élevé en valeur absolue que le plus élevé des 9 autres. En élevant au carré pour calculer l'erreur quadratique moyenne, ce rapport va se situer autour de 100. Il est donc certain que ce chiffre n'est pas réaliste et qu'il faudra une étude des valeurs aberrantes pour résoudre ce problème. Cependant on peut s'attendre à avoir des problèmes dans le cas stratifié avec une population cyclique. Ici les effets sont cycliques et dans plus de 34 % des échantillons, des effets de ligne ne sont pas estimés. Imaginons un cas où la grande majorité des effets de ligne non estimés tombe dans la région des effets faibles, combiné à la grande variabilité des données, il peut en résulter de graves problèmes de biais. Dans le cas aléatoire simple, il y a plus d'effets non estimés donc plus de chance de ne pas estimer des effets faibles et forts. De plus le fait que aussi bien des niveaux de ligne que de colonne peuvent ne pas être estimés augmente la chance de diversité dans les types d'effets non estimés ce qui est garant d'un biais plus faible. Pour le reste, les biais sont plus élevés que dans le cas précédent où le modèle n'était pas "faux" mais restent tout de même petits. Pour avoir une idée des estimateurs non

biaisés, nous avons fait des tests du genre  $H_0 : \text{biais} = 0$  contre  $H_1 : \text{biais} \neq 0$  avec le premier groupe d'échantillons de taille  $n = 500$ . Les résultats sont dans le tableau 3.7.

Comme on s'y attendait, seul le biais dans le cas cyclique avec échantillonnage stratifié

|          | AS     |        | ST     |           | NS     |        |
|----------|--------|--------|--------|-----------|--------|--------|
|          | Plan   | Modèle | Plan   | Modèle    | Plan   | Modèle |
| Normale  | 0.9827 | 0.8838 | 0.2993 | 0.2754    | 0.0744 | 0.0730 |
| Ar(1)    | 0.4891 | 0.3169 | 0.1724 | 0.0879    | 0.7324 | 0.7498 |
| Cyclique | 0.5267 | 0.1442 | 0.5873 | < 2.2e-16 | 0.5333 | 0.5488 |

TAB. 3.7 –  $p$ -value du test  $H_0 : \text{biais} = 0$  contre  $H_1 : \text{biais} \neq 0$

est biaisé. Tous les autres sont non biaisés avec une confiance de 95 %.

La racine carrée des erreurs quadratiques moyennes est donnée dans le tableau 3.8. L'estimation n'est plus aussi précise que précédemment avec des valeurs de l'ordre de 18 % à 67 %. La population AR(1) possède des CV estimés plus élevés, c'est une conséquence du fait que cette population est beaucoup plus variable que les deux autres. Ce qui est intéressant ici est de noter que les estimations par le modèle sont du même ordre que celles par le plan sauf dans le cas de la population cyclique avec échantillonnage stratifié. Donc en dépit du fait que le modèle linéaire n'est pas le modèle suivi par les données, l'estimation par le modèle est toujours appropriée quoique moins précise que celle par rapport au plan d'échantillonnage. Avec l'échantillonnage non standard, l'estimation par le modèle fait aussi bien que l'estimation par le plan. Cependant le cas de la population cyclique avec un échantillonnage stratifié nous rappelle la vigilance. L'échantillonnage non standard fait toujours mieux que les autres.

|          | AS    |        | ST    |        | NS    |        |
|----------|-------|--------|-------|--------|-------|--------|
|          | Plan  | Modèle | Plan  | Modèle | Plan  | Modèle |
| Normale  | 22.42 | 24.19  | 19.40 | 22.03  | 18.22 | 18.22  |
| Ar(1)    | 53.50 | 67.45  | 50.20 | 57.31  | 50.77 | 50.77  |
| Cyclique | 22.87 | 25.48  | 20.50 | 951.69 | 18.59 | 18.59  |

TAB. 3.8 – Moyenne des 10 estimations de CV en pourcentage quand le modèle est "faux".

En dehors du cas particulier de l'échantillonnage stratifié de la population cyclique, il semble que l'estimation par le modèle est moins précise mais reste assez proche de l'estimation par rapport au plan d'échantillonnage.

# Conclusion

L'essai a porté sur l'approche prédictive de l'échantillonnage. Nous nous sommes limité aux modèles linéaires généraux. Vu la vaste utilisation de ces modèles, cela n'est pas une trop grosse restriction.

Dans le cas des modèles avec variables auxiliaires continues, la théorie est assez simple. Le théorème général de la prédiction (Royall, 1976b) donne le meilleur prédicteur linéaire parmi tous les estimateurs sans biais. Ce meilleur prédicteur linéaire donné par le théorème est dans les types d'échantillonnage classique souvent identique à l'estimateur par le plan d'échantillonnage. Les échantillons équilibrés, au degré approprié, assurent le caractère non biaisé de nos prédicteurs. Cependant en pratique, il peut être très difficile de tirer des échantillons strictement équilibrés jusqu'au degré voulu et les études montrent que les différents estimateurs peuvent être assez sensibles au degré d'équilibrage atteint.

Dans le cas des variables auxiliaires qualitatives, il arrive que la matrice  $X'_s V_{ss}^{-1} X_s$  ne soit pas inversible. La théorie des inverses généralisés sera alors nécessaire pour résoudre le cas. Même si l'inverse généralisé n'est pas unique, il faut noter que l'estimateur optimum, lui est invariant c'est à dire unique. L'approche prédictive fonctionne donc dans le cas des modèles avec variables qualitatives même s'il y a sur-paramétrisation des niveaux des facteurs.

La dernière partie est consacrée à une simulation dans le cas particulier des modèles avec variables qualitatives (facteurs). Il s'agit d'un modèle linéaire avec deux facteurs et une seule observation par cellule. Les résultats montrent que tous les estimateurs selon le plan et le modèle sont non biaisés. Quand le modèle est "faux", l'estimateur stratifié avec la population cyclique est biaisé. L'estimation du  $CV$  montre que la méthode par le modèle est plus précise quand le modèle est correct. Quand le modèle est "faux", les estimations avec le plan sont meilleures toutefois celles avec le modèle restent proches sauf pour l'échantillonnage stratifié de la population cyclique. Avec l'échantillonnage non standard, dans tous les cas de figures, les estimations par le modèle sont équivalentes

à celles par le plan. Il semble que l'écart observé entre les méthodes prédictive et classique soit en grande partie dû aux effets non estimés. En effet, si nous considérons des populations avec des effets homogènes (modèle normal), plus il y a des effets non estimés plus le  $CV$  estimé est grand. Pour obtenir des estimations fiables et surtout plus précise par le modèle, il semble évident que des méthodes d'imputation efficaces pour estimer les effets manquants soient nécessaires.

Il ressort de cette étude que l'estimation par le modèle est une excellente alternative à celle par le plan quand le modèle suivi par les données est connu. Quand ce modèle n'est pas connu, cas plus réaliste, le fait de tirer des échantillons équilibrés jusqu'au degré nécessaire protège nos prédicteurs contre le biais et les rend très intéressants.

# Annexe A

## Données sur la population d'hôpitaux

| <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> |
|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|
| 1         | 10       | 57       | 2         | 14       | 64       | 3         | 15       | 41       | 4         | 15       | 76       | 5         | 16       | 35       |
| 6         | 18       | 56       | 7         | 19       | 42       | 8         | 19       | 48       | 9         | 20       | 23       | 10        | 20       | 90       |
| 11        | 20       | 22       | 12        | 24       | 170      | 13        | 24       | 120      | 14        | 25       | 36       | 15        | 25       | 49       |
| 16        | 25       | 14       | 17        | 25       | 92       | 18        | 25       | 64       | 19        | 25       | 103      | 20        | 26       | 100      |
| 21        | 26       | 98       | 22        | 27       | 89       | 23        | 28       | 85       | 24        | 29       | 109      | 25        | 30       | 79       |
| 26        | 30       | 79       | 27        | 32       | 52       | 28        | 32       | 125      | 29        | 34       | 64       | 30        | 35       | 100      |
| 31        | 35       | 75       | 32        | 37       | 108      | 33        | 38       | 95       | 34        | 38       | 78       | 35        | 39       | 153      |
| 36        | 40       | 124      | 37        | 40       | 87       | 38        | 40       | 121      | 39        | 41       | 213      | 40        | 43       | 141      |
| 41        | 43       | 141      | 42        | 47       | 174      | 43        | 48       | 81       | 44        | 49       | 173      | 45        | 50       | 260      |
| 46        | 50       | 186      | 47        | 50       | 296      | 48        | 50       | 87       | 49        | 50       | 229      | 50        | 50       | 194      |
| 51        | 56       | 115      | 52        | 57       | 220      | 53        | 57       | 247      | 54        | 59       | 297      | 55        | 61       | 308      |
| 56        | 61       | 58       | 57        | 62       | 91       | 58        | 62       | 182      | 59        | 63       | 242      | 60        | 63       | 222      |
| 61        | 64       | 240      | 62        | 64       | 225      | 63        | 65       | 239      | 64        | 65       | 231      | 65        | 67       | 255      |
| 66        | 67       | 321      | 67        | 67       | 215      | 68        | 68       | 259      | 69        | 69       | 233      | 70        | 69       | 253      |
| 71        | 70       | 209      | 72        | 70       | 216      | 73        | 70       | 315      | 74        | 70       | 233      | 75        | 70       | 258      |
| 76        | 70       | 244      | 77        | 73       | 200      | 78        | 74       | 297      | 79        | 80       | 297      | 80        | 80       | 301      |
| 81        | 81       | 266      | 82        | 86       | 270      | 83        | 88       | 310      | 84        | 90       | 326      | 85        | 91       | 120      |
| 86        | 95       | 243      | 87        | 95       | 243      | 88        | 96       | 377      | 89        | 96       | 228      | 90        | 98       | 308      |

---

*Nh* - Numéro de l'hôpital

*x* - Nombre de lits

*y* - Nombre de patients sortis

| <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> |
|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|
| 91        | 99       | 346      | 92        | 100      | 444      | 93        | 100      | 362      | 94        | 100      | 383      | 95        | 100      | 318      |
| 96        | 100      | 373      | 97        | 100      | 414      | 98        | 100      | 265      | 99        | 102      | 227      | 100       | 102      | 371      |
| 101       | 103      | 518      | 102       | 309      | 327      | 103       | 106      | 311      | 104       | 108      | 327      | 105       | 110      | 389      |
| 106       | 110      | 439      | 107       | 110      | 368      | 108       | 110      | 298      | 109       | 111      | 273      | 110       | 111      | 498      |
| 111       | 111      | 95       | 112       | 113      | 594      | 113       | 116      | 440      | 114       | 118      | 134      | 115       | 119      | 345      |
| 116       | 119      | 360      | 117       | 120      | 431      | 118       | 120      | 467      | 119       | 120      | 414      | 120       | 121      | 467      |
| 121       | 122      | 534      | 122       | 123      | 416      | 123       | 125      | 231      | 124       | 126      | 323      | 125       | 127      | 535      |
| 126       | 128      | 323      | 127       | 128      | 577      | 128       | 129      | 707      | 129       | 130      | 426      | 130       | 134      | 381      |
| 131       | 135      | 411      | 132       | 135      | 362      | 133       | 137      | 505      | 134       | 138      | 592      | 135       | 139      | 244      |
| 136       | 141      | 355      | 137       | 142      | 322      | 138       | 143      | 384      | 139       | 144      | 470      | 140       | 145      | 828      |
| 141       | 145      | 475      | 142       | 145      | 337      | 143       | 145      | 283      | 144       | 150      | 470      | 145       | 150      | 621      |
| 146       | 151      | 376      | 147       | 151      | 543      | 148       | 152      | 538      | 149       | 154      | 486      | 150       | 155      | 467      |
| 151       | 156      | 778      | 152       | 159      | 487      | 153       | 160      | 637      | 154       | 160      | 590      | 155       | 160      | 402      |
| 156       | 161      | 611      | 157       | 163      | 690      | 158       | 165      | 360      | 159       | 169      | 662      | 160       | 170      | 689      |
| 161       | 170      | 665      | 162       | 175      | 592      | 163       | 178      | 446      | 164       | 180      | 84       | 165       | 180      | 479      |
| 166       | 180      | 531      | 167       | 181      | 573      | 168       | 184      | 481      | 169       | 184      | 652      | 170       | 185      | 635      |
| 171       | 187      | 1011     | 172       | 188      | 713      | 173       | 192      | 625      | 174       | 193      | 504      | 175       | 195      | 744      |
| 176       | 196      | 586      | 177       | 200      | 695      | 178       | 204      | 697      | 179       | 204      | 670      | 180       | 205      | 622      |
| 181       | 206      | 703      | 182       | 207      | 814      | 183       | 207      | 726      | 184       | 210      | 670      | 185       | 214      | 918      |
| 186       | 214      | 726      | 187       | 224      | 590      | 188       | 224      | 587      | 189       | 224      | 558      | 190       | 225      | 1186     |
| 191       | 227      | 410      | 192       | 227      | 732      | 193       | 228      | 955      | 194       | 229      | 1175     | 195       | 229      | 439      |
| 196       | 231      | 931      | 197       | 233      | 684      | 198       | 235      | 669      | 199       | 235      | 629      | 200       | 235      | 925      |
| 201       | 241      | 610      | 202       | 242      | 601      | 203       | 244      | 858      | 204       | 244      | 490      | 205       | 247      | 1084     |
| 206       | 247      | 1028     | 207       | 248      | 928      | 208       | 252      | 810      | 209       | 252      | 995      | 210       | 254      | 956      |
| 211       | 255      | 1160     | 212       | 256      | 705      | 213       | 257      | 974      | 214       | 260      | 1076     | 215       | 261      | 788      |
| 216       | 261      | 795      | 217       | 263      | 811      | 218       | 264      | 1009     | 219       | 265      | 609      | 220       | 268      | 1106     |
| 221       | 268      | 773      | 222       | 269      | 884      | 223       | 269      | 887      | 224       | 270      | 951      | 225       | 273      | 956      |
| 226       | 275      | 1201     | 227       | 275      | 1063     | 228       | 275      | 632      | 229       | 276      | 852      | 230       | 279      | 754      |
| 231       | 279      | 861      | 232       | 282      | 767      | 233       | 284      | 456      | 234       | 285      | 1007     | 235       | 286      | 941      |
| 236       | 287      | 1097     | 237       | 289      | 233      | 238       | 291      | 824      | 239       | 295      | 764      | 240       | 297      | 842      |
| 241       | 297      | 539      | 242       | 300      | 778      | 243       | 300      | 557      | 244       | 302      | 958      | 245       | 303      | 715      |
| 246       | 304      | 1036     | 247       | 307      | 1153     | 248       | 307      | 855      | 249       | 307      | 935      | 250       | 308      | 1031     |

---

*Nh* - Numéro de l'hôpital

*x* - Nombre de lits

*y* - Nombre de patients sortis

| <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> | <i>Nh</i> | <i>x</i> | <i>y</i> |
|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|-----------|----------|----------|
| 251       | 309      | 985      | 252       | 310      | 1042     | 253       | 310      | 1226     | 254       | 312      | 912      | 255       | 313      | 1016     |
| 256       | 318      | 944      | 257       | 322      | 876      | 258       | 324      | 1232     | 259       | 325      | 1049     | 260       | 327      | 1210     |
| 261       | 327      | 946      | 262       | 330      | 471      | 263       | 330      | 872      | 264       | 332      | 1400     | 265       | 338      | 1425     |
| 266       | 339      | 885      | 267       | 340      | 1133     | 268       | 340      | 1097     | 269       | 347      | 1166     | 270       | 347      | 906      |
| 271       | 348      | 1219     | 272       | 350      | 1173     | 273       | 352      | 1098     | 274       | 354      | 876      | 275       | 357      | 915      |
| 276       | 358      | 976      | 277       | 359      | 1029     | 278       | 361      | 889      | 279       | 365      | 966      | 280       | 365      | 956      |
| 281       | 365      | 1766     | 282       | 366      | 1225     | 283       | 367      | 1453     | 284       | 368      | 1413     | 285       | 370      | 1156     |
| 286       | 373      | 787      | 287       | 374      | 1137     | 288       | 374      | 1231     | 289       | 378      | 896      | 290       | 380      | 1009     |
| 291       | 385      | 1150     | 292       | 386      | 1272     | 293       | 390      | 1373     | 294       | 391      | 1389     | 295       | 393      | 926      |
| 296       | 394      | 1060     | 297       | 400      | 2190     | 298       | 400      | 1219     | 299       | 400      | 1095     | 300       | 401      | 1634     |
| 301       | 408      | 1719     | 302       | 411      | 989      | 303       | 411      | 808      | 304       | 417      | 1369     | 305       | 418      | 1040     |
| 306       | 419      | 1315     | 307       | 422      | 1089     | 308       | 425      | 1347     | 309       | 437      | 1632     | 310       | 437      | 1346     |
| 311       | 438      | 1370     | 312       | 445      | 1105     | 313       | 450      | 1705     | 314       | 451      | 1584     | 315       | 461      | 1948     |
| 316       | 463      | 1617     | 317       | 467      | 1665     | 318       | 469      | 1012     | 319       | 470      | 1322     | 320       | 472      | 1239     |
| 321       | 474      | 1258     | 322       | 478      | 1835     | 323       | 479      | 1534     | 324       | 480      | 1149     | 325       | 490      | 1390     |
| 326       | 492      | 1126     | 327       | 493      | 1355     | 328       | 496      | 1301     | 329       | 498      | 1657     | 330       | 500      | 1785     |
| 331       | 500      | 1744     | 332       | 505      | 1669     | 333       | 506      | 1527     | 334       | 509      | 2031     | 335       | 510      | 2051     |
| 336       | 517      | 834      | 337       | 523      | 1232     | 338       | 524      | 1350     | 339       | 524      | 1805     | 340       | 530      | 1420     |
| 341       | 534      | 2034     | 342       | 536      | 1418     | 343       | 538      | 1522     | 344       | 540      | 1386     | 345       | 541      | 1376     |
| 346       | 543      | 1093     | 347       | 543      | 1780     | 348       | 549      | 1547     | 349       | 550      | 986      | 350       | 550      | 1287     |
| 351       | 551      | 1645     | 352       | 556      | 1478     | 353       | 558      | 1152     | 354       | 562      | 2116     | 355       | 566      | 1828     |
| 356       | 573      | 1789     | 357       | 577      | 1509     | 358       | 579      | 1415     | 359       | 583      | 1583     | 360       | 584      | 1326     |
| 361       | 591      | 999      | 362       | 592      | 1648     | 363       | 600      | 2154     | 364       | 606      | 1785     | 365       | 606      | 1218     |
| 366       | 613      | 1463     | 367       | 625      | 2240     | 368       | 631      | 1684     | 369       | 635      | 1606     | 370       | 650      | 1620     |
| 371       | 652      | 2150     | 372       | 658      | 1376     | 373       | 670      | 1707     | 374       | 684      | 1504     | 375       | 712      | 1893     |
| 376       | 712      | 2089     | 377       | 719      | 2058     | 378       | 760      | 1283     | 379       | 774      | 2844     | 380       | 785      | 2171     |
| 381       | 816      | 1239     | 382       | 817      | 1706     | 383       | 829      | 2766     | 384       | 830      | 1715     | 385       | 838      | 2135     |
| 386       | 857      | 1624     | 387       | 860      | 2818     | 388       | 904      | 2700     | 389       | 918      | 1678     | 390       | 936      | 1394     |
| 391       | 937      | 1894     | 392       | 957      | 1765     | 393       | 986      | 2268     |           |          |          |           |          |          |

TAB. A.1 – Données de la population de  $N=393$  hôpitaux

---

*Nh* - Numéro de l'hôpital

*x* - Nombre de lits

*y* - Nombre de patients sortis

# Annexe B

## Construction d'un g-inverse par la méthode de décomposition en valeur singulière

Un inverse généralisé d'une matrice peut être construit en utilisant la décomposition en valeur singulière ou en anglais *singular value decomposition* (SVD) de la matrice. La méthode est décrite dans ce qui suit.

Une matrice  $A$  de valeurs réelles de dimension  $p \times q$  peut être décomposée comme un produit d'une matrice  $U$  orthogonale de dimension  $p \times p$ , une matrice  $D$  diagonale de dimension  $q \times q$  et une autre matrice  $V$  orthogonale de dimension  $q \times q$  :

$$A = UDV'. \quad (\text{B.1})$$

Les matrices  $U$  et  $V$  sont orthogonales c'est à dire que  $U'U = V'V = I_q$  où  $I_q$  est la matrice identité de dimension  $q \times q$ . Les éléments de  $D$  sont appelés les valeurs singulières de  $A$ . Les carrés des valeurs singulières de  $A$  sont les valeurs propres de la matrice  $A'A$ . L'Expression (B.1) est la décomposition en valeur singulière (SVD) de la matrice  $A$ . Si  $A$  est symétrique alors  $U = V$  et

$$A^{1/2} = UD^{1/2}V'. \quad (\text{B.2})$$

où  $D^{1/2}$  est la matrice diagonale de la racine carrée des éléments de  $D$ .

En utilisant le SVD, un g-inverse de  $A$  est  $G = VD^{-1}U'$ . En effet nous avons :

$$\begin{aligned}AGA &= UDV'(VD^{-1}U')UDV' \\ &= UDV' \\ &= A.\end{aligned}$$

Quand au moins un des éléments de  $D$  est nul c'est à dire la matrice  $A'A$  a au moins une valeur propre égale à zéro alors  $G$  est modifié comme suit. Toute paire de ligne et colonne correspondante à un zéro sur la diagonale de  $D$  est supprimée soit  $D^*$  la matrice résultante. De même les lignes et colonnes de  $U$  et  $V$  correspondantes sont supprimées pour obtenir  $U^*$  et  $V^*$ . Le g-inverse de  $A$  est alors  $G = V^*D^{*-1}U^{*'}.$

Le lemme qui suit est utile pour l'étude des modèles avec covariables.  $M$  est une matrice définie comme suit :

$$M = \begin{pmatrix} X' \\ Z' \end{pmatrix} \begin{pmatrix} X & Z \end{pmatrix} \equiv \begin{pmatrix} A & B \\ B' & D \end{pmatrix}$$

où  $X$  est la matrice des variables qualitatives et  $Z$  celle des variables quantitatives.

**Lemme 6.** *Un inverse généralisé de  $M$  est :*

$$\begin{aligned} M^- &= \begin{pmatrix} A^- + A^-BQ^-B'A^- & -A^-BQ^- \\ -Q^-B'A^- & Q^- \end{pmatrix} \\ &= \begin{pmatrix} A^- & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -A^-B \\ I \end{pmatrix} Q^- \begin{pmatrix} -B'A^- & I \end{pmatrix}, \end{aligned}$$

où  $A^-$  est un g-inverse de  $A$ ,  $Q = D - B'A^-B$  et  $Q^-$  est un g-inverse de  $Q$ .

# Annexe C

## Programmes R pour la simulation

### C.1 La fonction population()

Cette fonction permet de générer une population avec deux variables auxiliaires.

- mu est la moyenne globale et X1 représente la variable ligne alors que X2 est la colonne.
- seed est la racine aléatoire, elle permet d’avoir les mêmes échantillons à chaque reprise.
- b est le coefficient de la série autoregressive et var\_erreur est la variance des résidus du modèle.

```
population <- fonction (mu=5, var_X1=1, var_X2=1.25, n_X1=30, n_X2=30, b=0.9,
var_erreur=.5, seed=312557671){
# fixer la racine aléatoire
set.seed(seed)

# Genere les effets selon la loi normale
X1_norm <- rep(0, n_X1)
X2_norm <- rep(0, n_X2)
X1_norm <- rnorm(n_X1, 0, sqrt(var_X1))
X1_norm <- X1_norm - rep(mean(X1_norm), n_X1)
X2_norm <- rnorm(n_X2, 0, sqrt(var_X2))
X2_norm <- X2_norm - rep(mean(X2_norm), n_X2)

# Genere les effets selon la loi autoregressif d'ordre 1 AR(1)
X1_ar <- rep(0, n_X1)
```

```

X2_ar <- rep(0, n_X2)
std_alphai <- sqrt(var_X1 / (1-b^2))
std_betai <- sqrt(var_X2 / (1-b^2))
X1_ar[1] <- rnorm(1, 0, sd=sqrt(std_alphai))
X2_ar[1] <- rnorm(1, 0, sd=sqrt(std_betai))
for( i in 2 :n_X1)
X1_ar[i] <- 0.9*X1_ar[i-1] + rnorm(1, 0, sd=sqrt(var_X1))
for( i in 2 :n_X2)
X2_ar[i] <- 0.9*X2_ar[i-1] + rnorm(1, 0, sd=sqrt(var_X2))

X1_ar <- X1_ar - rep(mean(X1_ar),n_X1)
X2_ar <- X2_ar - rep(mean(X2_ar),n_X2)

# Genere les effets selon la loi périodique sinus
X1_cycle <- rep(0, n_X1)
X2_cycle <- rep(0, n_X2)
X1_cycle[1] <- rnorm(1, 0, sqrt(var_X1))
X2_cycle[1] <- rnorm(1, 0, sqrt(var_X2))
for( i in 2 :n_X1)
X1_cycle[i] <- sin(2*pi*i/n_X1) + rnorm(1, 0, sqrt(var_X1))
for( j in 2 :n_X2)
X2_cycle[j] <- sin(2*pi*j/n_X2) + rnorm(1, 0, sqrt(var_X2))

X1_cycle <- X1_cycle - rep(mean(X1_cycle), n_X1)
X2_cycle <- X2_cycle - rep(mean(X2_cycle), n_X2)

# Genere les realisations de la variable réponse selon un modèle à effets additifs
# et un modèle à effets multiplicatifs
Yad_norm <- matrix(0, n_X1, n_X2)
Yad_ar <- matrix(0, n_X1, n_X2)
Yad_cycle <- matrix(0, n_X1, n_X2)
Ymul_norm <- matrix(0, n_X1, n_X2)
Ymul_ar <- matrix(0, n_X1, n_X2)
Ymul_cycle <- matrix(0, n_X1, n_X2)
for( i in 1 :n_X1)
for( j in 1 :n_X2){
Yad_norm[i,j] = mu + X1_norm[i] + X2_norm[j] + rnorm(1, 0, sqrt(var_erreur))
Yad_ar[i,j] = mu + X1_ar[i] + X2_ar[j] + rnorm(1, 0, sqrt(var_erreur))
Yad_cycle[i,j] = mu + X1_cycle[i] + X2_cycle[j] + rnorm(1, 0, sqrt(var_erreur))
Ymul_norm[i,j] = exp(mu + X1_norm[i] + X2_norm[j] + rnorm(1, 0, sqrt(var_erreur)))
Ymul_ar[i,j] = exp(mu + X1_ar[i] + X2_ar[j] + rnorm(1, 0, sqrt(var_erreur)))
Ymul_cycle[i,j] = exp(mu + X1_cycle[i] + X2_cycle[j] + rnorm(1, 0, sqrt(var_erreur)))
}

```

```

# Calcul de la variance échantillonnale
Vad_norm <- var(as.vector(Yad_norm))
Vad_ar <- var(as.vector(Yad_ar))
Vad_cycle <- var(as.vector(Yad_cycle))
Vmul_norm <- var(as.vector(Ymul_norm))
Vmul_ar <- var(as.vector(Ymul_ar))
Vmul_cycle <- var(as.vector(Ymul_cycle))

# Calcul de coefficient de variation
CVad_norm <- sd(as.vector(Yad_norm))/mean(Yad_norm)
CVad_ar <- sd(as.vector(Yad_ar))/mean(Yad_ar)
CVad_cycle <- sd(as.vector(Yad_cycle))/mean(Yad_cycle)
CVmul_norm <- sd(as.vector(Ymul_norm))/mean(Ymul_norm)
CVmul_ar <- sd(as.vector(Ymul_ar))/mean(Ymul_ar)
CVmul_cycle <- sd(as.vector(Ymul_cycle))/mean(Ymul_cycle)

obj <- list(X1_norm=X1_norm, X2_norm=X2_norm, X1_ar=X1_ar, X2_ar=X2_ar,
X1_cycle=X1_cycle, X2_cycle=X2_cycle, std_alphai=std_alphai,
std_betai=std_betai, mu=mu, Yad_norm=Yad_norm, Yad_ar=Yad_ar,
Yad_cycle=Yad_cycle, Ymul_norm=Ymul_norm, Ymul_ar=Ymul_ar,
Ymul_cycle=Ymul_cycle, CVad_norm=CVad_norm, CVad_ar=CVad_ar,
CVad_cycle=CVad_cycle, CVmul_norm=CVmul_norm, CVmul_ar=CVmul_ar,
CVmul_cycle=CVmul_cycle, Vad_norm=Vad_norm, Vad_ar=Vad_ar,
Vad_cycle=Vad_cycle, Vmul_norm=Vmul_norm, Vmul_ar=Vmul_ar,
Vmul_cycle=Vmul_cycle, n_X1=n_X1, n_X2=n_X2, N=n_X1*n_X2)
}

```

## C.2 La fonction tirage()

- Cette fonction fait le tirage selon 3 méthodes : aléatoire simple, stratifié et un dernier tirage non standard.
- pop est l'objet retourné par la fonction population(), il contient toute l'information sur la population.
- n est la taille de l'échantillon et doit être un multiple de la longueur de strates (colonnes) et du nombre de ligne.
- Si plan="vrai" alors le modèle est "vrai" sinon le modèle est "faux".

```

tirage <- fonction(pop, n=120, plan="vrai") {
# Initialisation
Y_norm <- matrix(NA, pop$n_X1, pop$n_X2)

```

```

Y_ar <- matrix(NA, pop$n_X1, pop$n_X2)
Y_cycle <- matrix(NA, pop$n_X1, pop$n_X2)
if(plan=="vrai") {
  Y_norm <- pop$Yad_norm
  Y_ar <- pop$Yad_ar
  Y_cycle <- pop$Yad_cycle
} else {
  Y_norm <- pop$Ymul_norm
  Y_ar <- pop$Ymul_ar
  Y_cycle <- pop$Ymul_cycle
}
Donn_norm <- cbind(as.vector(Y_norm), gl(30,1,900), gl(30,30))
Donn_ar <- cbind(as.vector(Y_ar), gl(30,1,900), gl(30,30))
Donn_cycle <- cbind(as.vector(Y_cycle), gl(30,1,900), gl(30,30))

# Tirage aléatoire simple
indiceAS <- sample(1 :pop$N, n)
Ys_normAS <- rep(NA, n)
Ys_arAS <- rep(NA, n)
Ys_cycleAS <- rep(NA, n)
Ys_normAS <- Donn_norm[indiceAS,]
Ys_arAS <- Donn_ar[indiceAS,]
Ys_cycleAS <- Donn_cycle[indiceAS,]

# Tirage stratifié
indiceST_X1 <- sample(1 :pop$n_X1, n/pop$n_X1)
for (i in 1 :(pop$n_X1-1))
  indiceST_X1 <- c(indiceST_X1, sample(1 :pop$n_X1, n/pop$n_X1))
indiceST_X2 <- gl(pop$n_X2, n/pop$n_X2)
indiceST <- cbind(indiceST_X1, indiceST_X2)

Ys_normST <- rep(NA, n)
Ys_arST <- rep(NA, n)
Ys_cycleST <- rep(NA, n)
Ys_normST <- Donn_norm[indiceST[,1]+pop$n_X1*(indiceST[,2]-1),]
Ys_arST <- Donn_ar[indiceST[,1]+pop$n_X1*(indiceST[,2]-1),]
Ys_cycleST <- Donn_cycle[indiceST[,1]+pop$n_X1*(indiceST[,2]-1),]

# Tirage non standard
indiceNS_X1 <- sample(gl(pop$n_X1, n/pop$n_X1))
indiceNS_X2 <- sample(gl(pop$n_X2, n/pop$n_X2))
indiceNS <- cbind(indiceNS_X1, indiceNS_X2)

```



```
MoyMod_cycleST <- rep(NA, nrep)
Ysmod_normNS <- array(NA, dim=c(Ys$n, 3, nrep))
Ysmod_arNS <- array(NA, dim=c(Ys$n, 3, nrep))
Ysmod_cycleNS <- array(NA, dim=c(Ys$n, 3, nrep))
MoyMod_normNS <- rep(NA, nrep)
MoyMod_arNS <- rep(NA, nrep)
MoyMod_cycleNS <- rep(NA, nrep)
```

```
MoyPlan_normAS <- rep(NA, nrep)
MoyPlan_arAS <- rep(NA, nrep)
MoyPlan_cycleAS <- rep(NA, nrep)
MoyPlan_normST <- rep(NA, nrep)
MoyPlan_arST <- rep(NA, nrep)
MoyPlan_cycleST <- rep(NA, nrep)
MoyPlan_normNS <- rep(NA, nrep)
MoyPlan_arNS <- rep(NA, nrep)
MoyPlan_cycleNS <- rep(NA, nrep)
ECPlan_normAS <- rep(NA, nrep)
ECPlan_normST <- rep(NA, nrep)
ECPlan_normNS <- rep(NA, nrep)
ECPlan_arAS <- rep(NA, nrep)
ECPlan_arST <- rep(NA, nrep)
ECPlan_arNS <- rep(NA, nrep)
ECPlan_cycleAS <- rep(NA, nrep)
ECPlan_cycleST <- rep(NA, nrep)
ECPlan_cycleNS <- rep(NA, nrep)
```

```
MoyMod_normAS <- rep(NA, nrep)
MoyMod_arAS <- rep(NA, nrep)
MoyMod_cycleAS <- rep(NA, nrep)
MoyMod_normST <- rep(NA, nrep)
MoyMod_arST <- rep(NA, nrep)
MoyMod_cycleST <- rep(NA, nrep)
MoyMod_normNS <- rep(NA, nrep)
MoyMod_arNS <- rep(NA, nrep)
MoyMod_cycleNS <- rep(NA, nrep)
ECmod_normAS <- rep(NA, nrep)
ECmod_normST <- rep(NA, nrep)
ECmod_normNS <- rep(NA, nrep)
ECmod_arAS <- rep(NA, nrep)
ECmod_arST <- rep(NA, nrep)
ECmod_arNS <- rep(NA, nrep)
ECmod_cycleAS <- rep(NA, nrep)
```

```

ECmod_cycleST <- rep(NA, nrep)
ECmod_cycleNS <- rep(NA, nrep)

PropCompl_AS <- 0
PropCompl_ST <- 0
PropCompl_NS <- 0

# cette condition permet de faire les inference selon que le modèle est correct ou faux
if (plan == "vrai") {
# Totaux dans la population entière
Y_norm <- pop$Yad_norm
Y_ar <- pop$Yad_ar
Y_cycle <- pop$Yad_cycle
Moy_norm <- mean(Y_norm)
Moy_ar <- mean(Y_ar)
Moy_cycle <- mean(Y_cycle)
} else {
Y_norm <- pop$Ymul_norm
Y_ar <- pop$Ymul_ar
Y_cycle <- pop$Ymul_cycle
Moy_norm <- mean(Y_norm)
Moy_ar <- mean(Y_ar)
Moy_cycle <- mean(Y_cycle)
}

# la boucle permet de faire plusieurs estimations dans le but de calculer des variances etc...
for (k in 1 :nrep) {
cat (k,"
n")
# Affectation initiale
Ys <- tirage(pop=pop, plan=plan)
Ysmod_normAS[,k] <- Ys$Ys_normAS
Ysmod_arAS[,k] <- Ys$Ys_arAS
Ysmod_cycleAS[,k] <- Ys$Ys_cycleAS
Ysmod_normST[,k] <- Ys$Ys_normST
Ysmod_arST[,k] <- Ys$Ys_arST
Ysmod_cycleST[,k] <- Ys$Ys_cycleST
Ysmod_normNS[,k] <- Ys$Ys_normNS
Ysmod_arNS[,k] <- Ys$Ys_arNS
Ysmod_cycleNS[,k] <- Ys$Ys_cycleNS

S1_AS <- sort(unique(Ysmod_normAS[,2,k])) NX1reel_AS <- length(S1_AS)
S2_AS <- sort(unique(Ysmod_normAS[,3,k])) NX2reel_AS <- length(S2_AS)

```

```
S1_ST <- sort(unique(Ysmod_normST[,2,k])) NX1reel_ST <- length(S1_ST)
S2_ST <- sort(unique(Ysmod_normST[,3,k])) NX2reel_ST <- length(S2_ST)
```

```
S1_NS <- sort(unique(Ysmod_normNS[,2,k])) NX1reel_NS <- length(S1_NS)
S2_NS <- sort(unique(Ysmod_normNS[,3,k])) NX2reel_NS <- length(S2_NS)
```

```
Modele_normAS <- lm (Ysmod_normAS[,1,k] factor(Ysmod_normAS[,2,k])
+ factor(Ysmod_normAS[,3,k]))
Coef_normAS <- Modele_normAS$coefficient
Modele_normST <- lm (Ysmod_normST[,1,k] factor(Ysmod_normST[,2,k])
+ factor(Ysmod_normST[,3,k]))
Coef_normST <- Modele_normST$coefficient
YsmodU_normNS <- unique(Ysmod_normNS)
Modele_normNS <- lm (YsmodU_normNS[,1,k] factor(YsmodU_normNS[,2,k])
+ factor(YsmodU_normNS[,3,k]))
Coef_normNS <- Modele_normNS$coefficient
```

```
Modele_arAS <- lm (Ysmod_arAS[,1,k] factor(Ysmod_arAS[,2,k])
+ factor(Ysmod_arAS[,3,k]))
Coef_arAS <- Modele_arAS$coefficient
Modele_arST <- lm (Ysmod_arST[,1,k] factor(Ysmod_arST[,2,k])
+ factor(Ysmod_arST[,3,k]))
Coef_arST <- Modele_arST$coefficient
YsmodU_arNS <- unique(Ysmod_arNS)
Modele_arNS <- lm (YsmodU_arNS[,1,k] factor(YsmodU_arNS[,2,k])
+ factor(YsmodU_arNS[,3,k]))
Coef_arNS <- Modele_arNS$coefficient
```

```
Modele_cycleAS <- lm (Ysmod_cycleAS[,1,k] factor(Ysmod_cycleAS[,2,k])
+ factor(Ysmod_cycleAS[,3,k]))
Coef_cycleAS <- Modele_cycleAS$coefficient
Modele_cycleST <- lm (Ysmod_cycleST[,1,k] factor(Ysmod_cycleST[,2,k])
+ factor(Ysmod_cycleST[,3,k]))
Coef_cycleST <- Modele_arST$coefficient
YsmodU_cycleNS <- unique(Ysmod_cycleNS)
Modele_cycleNS <- lm (YsmodU_cycleNS[,1,k] factor(YsmodU_cycleNS[,2,k])
+ factor(YsmodU_cycleNS[,3,k]))
Coef_cycleNS <- Modele_cycleNS$coefficient
```

```
Ymod_normAS <- matrix(NA, pop$n_X1, pop$n_X2)
Ymod_arAS <- matrix(NA, pop$n_X1, pop$n_X2)
Ymod_cycleAS <- matrix(NA, pop$n_X1, pop$n_X2)
```

```

Ymod_normST <- matrix(NA, pop$n_X1, pop$n_X2)
Ymod_arST <- matrix(NA, pop$n_X1, pop$n_X2)
Ymod_cycleST <- matrix(NA, pop$n_X1, pop$n_X2)
Ymod_normNS <- matrix(NA, pop$n_X1, pop$n_X2)
Ymod_arNS <- matrix(NA, pop$n_X1, pop$n_X2)
Ymod_cycleNS <- matrix(NA, pop$n_X1, pop$n_X2)

# Inference par le modèle avec l'échantillonnage aléatoire simple
for (i in 1 :NX1reel_AS)
for (j in 1 :NX2reel_AS) {
if ((i==1) & (j==1)) {
Ymod_normAS[S1_AS[i],S2_AS[j]] <- Coef_normAS[1]
Ymod_arAS[S1_AS[i],S2_AS[j]] <- Coef_arAS[1]
Ymod_cycleAS[S1_AS[i],S2_AS[j]] <- Coef_cycleAS[1]
} else if (i==1) {
Ymod_normAS[S1_AS[i],S2_AS[j]] <- Coef_normAS[1] + Coef_normAS[NX1reel_AS+j-1]
Ymod_arAS[S1_AS[i],S2_AS[j]] <- Coef_arAS[1] + Coef_arAS[NX1reel_AS+j-1]
Ymod_cycleAS[S1_AS[i],S2_AS[j]] <- Coef_cycleAS[1] + Coef_cycleAS[NX1reel_AS+j-1]
} else if (j==1) {
Ymod_normAS[S1_AS[i],S2_AS[j]] <- Coef_normAS[1] + Coef_normAS[i]
Ymod_arAS[S1_AS[i],S2_AS[j]] <- Coef_arAS[1] + Coef_arAS[i]
Ymod_cycleAS[S1_AS[i],S2_AS[j]] <- Coef_cycleAS[1] + Coef_cycleAS[i]
} else {
Ymod_normAS[S1_AS[i],S2_AS[j]] <- Coef_normAS[1] + Coef_normAS[i]
+ Coef_normAS[NX1reel_AS+j-1]
Ymod_arAS[S1_AS[i],S2_AS[j]] <- Coef_arAS[1] + Coef_arAS[i]
+ Coef_arAS[NX1reel_AS+j-1]
Ymod_cycleAS[S1_AS[i],S2_AS[j]] <- Coef_cycleAS[1] + Coef_cycleAS[i]
+ Coef_cycleAS[NX1reel_AS+j-1]
}
}

if ((NX1reel_AS==pop$n_X1) & (NX2reel_AS==pop$n_X2))
PropCompl_AS <- PropCompl_AS + 1

MoyMod_normAS[k] <- mean(Ymod_normAS, na.rm=TRUE)
MoyMod_arAS[k] <- mean(Ymod_arAS, na.rm=TRUE)
MoyMod_cycleAS[k] <- mean(Ymod_cycleAS, na.rm=TRUE)

ECmod_normAS[k] <- MoyMod_normAS[k] - Moy_norm
ECmod_arAS[k] <- MoyMod_arAS[k] - Moy_ar
ECmod_cycleAS[k] <- MoyMod_cycleAS[k] - Moy_cycle

```

```

# Inference par le modèle avec l'échantillonnage stratifié
for (i in 1 :NX1reel.ST)
for (j in 1 :NX2reel.ST) {
if ((i==1) & (j==1)) {
Ymod_normST[S1.ST[i],S2.ST[j]] <- Coef_normST[1]
Ymod_arST[S1.ST[i],S2.ST[j]] <- Coef_arST[1]
Ymod_cycleST[S1.ST[i],S2.ST[j]] <- Coef_cycleST[1]
} else if (i==1) {
Ymod_normST[S1.ST[i],S2.ST[j]] <- Coef_normST[1] + Coef_normST[NX1reel.ST+j-1]
Ymod_arST[S1.ST[i],S2.ST[j]] <- Coef_arST[1] + Coef_arST[NX1reel.ST+j-1]
Ymod_cycleST[S1.ST[i],S2.ST[j]] <- Coef_cycleST[1] + Coef_cycleST[NX1reel.ST+j-1]
} else if (j==1) {
Ymod_normST[S1.ST[i],S2.ST[j]] <- Coef_normST[1] + Coef_normST[i]
Ymod_arST[S1.ST[i],S2.ST[j]] <- Coef_arST[1] + Coef_arST[i]
Ymod_cycleST[S1.ST[i],S2.ST[j]] <- Coef_cycleST[1] + Coef_cycleST[i]
} else {
Ymod_normST[S1.ST[i],S2.ST[j]] <- Coef_normST[1] + Coef_normST[i]
+ Coef_normST[NX1reel.ST+j-1]
Ymod_arST[S1.ST[i],S2.ST[j]] <- Coef_arST[1] + Coef_arST[i]
+ Coef_arST[NX1reel.ST+j-1]
Ymod_cycleST[S1.ST[i],S2.ST[j]] <- Coef_cycleST[1] + Coef_cycleST[i]
+ Coef_cycleST[NX1reel.ST+j-1]
}
}
}

if ((NX1reel.ST==pop$n.X1) & (NX2reel.ST==pop$n.X2))
PropCompl.ST <- PropCompl.ST + 1

MoyMod_normST[k] <- mean(Ymod_normST, na.rm=TRUE)
MoyMod_arST[k] <- mean(Ymod_arST, na.rm=TRUE)
MoyMod_cycleST[k] <- mean(Ymod_cycleST, na.rm=TRUE)

ECmod_normST[k] <- MoyMod_normST[k] - Moy_norm
ECmod_arST[k] <- MoyMod_arST[k] - Moy_ar
ECmod_cycleST[k] <- MoyMod_cycleST[k] - Moy_cycle

# Inference par le modèle avec l'échantillonnage non standard
for (i in 1 :NX1reel.NS)
for (j in 1 :NX2reel.NS) {
if ((i==1) & (j==1)) {
Ymod_normNS[S1.NS[i], S2.NS[j]] <- Coef_normNS[1]
Ymod_arNS[S1.NS[i], S2.NS[j]] <- Coef_arNS[1]
Ymod_cycleNS[S1.NS[i], S2.NS[j]] <- Coef_cycleNS[1]
}
}
}

```

```

} else if (i==1) {
Ymod_normNS[S1_NS[i], S2_NS[j]] <- Coef_normNS[1] + Coef_normNS[NX1reel_NS+j-1]
Ymod_arNS[S1_NS[i], S2_NS[j]] <- Coef_arNS[1] + Coef_arNS[NX1reel_NS+j-1]
Ymod_cycleNS[S1_NS[i], S2_NS[j]] <- Coef_cycleNS[1] + Coef_cycleNS[NX1reel_NS+j-1]
} else if (j==1) {
Ymod_normNS[S1_NS[i], S2_NS[j]] <- Coef_normNS[1] + Coef_normNS[i]
Ymod_arNS[S1_NS[i], S2_NS[j]] <- Coef_arNS[1] + Coef_arNS[i]
Ymod_cycleNS[S1_NS[i], S2_NS[j]] <- Coef_cycleNS[1] + Coef_cycleNS[i]
} else {
Ymod_normNS[S1_NS[i], S2_NS[j]] <- Coef_normNS[1] + Coef_normNS[i]
+ Coef_normNS[NX1reel_NS+j-1]
Ymod_arNS[S1_NS[i], S2_NS[j]] <- Coef_arNS[1] + Coef_arNS[i]
+ Coef_arNS[NX1reel_NS+j-1]
Ymod_cycleNS[S1_NS[i], S2_NS[j]] <- Coef_cycleNS[1] + Coef_cycleNS[i]
+ Coef_cycleNS[NX1reel_NS+j-1]
}
}
}

```

```

if ((NX1reel_NS==pop$n_X1) & (NX2reel_NS==pop$n_X2))
PropCompl_NS <- PropCompl_NS + 1

```

```

MoyMod_normNS[k] <- mean(Ymod_normNS, na.rm=TRUE)
MoyMod_arNS[k] <- mean(Ymod_arNS, na.rm=TRUE)
MoyMod_cycleNS[k] <- mean(Ymod_cycleNS, na.rm=TRUE)

```

```

ECmod_normNS[k] <- MoyMod_normNS[k] - Moy_norm
ECmod_arNS[k] <- MoyMod_arNS[k] - Moy_ar
ECmod_cycleNS[k] <- MoyMod_cycleNS[k] - Moy_cycle

```

```

# Inference par le plan

```

```

MoyPlan_normAS[k] <- sum(Ys$Ys_normAS[,1])/Ys$n
ECPlan_normAS[k] <- MoyPlan_normAS[k] - Moy_norm
MoyPlan_arAS[k] <- sum(Ys$Ys_arAS[,1])/Ys$n
ECPlan_arAS[k] <- MoyPlan_arAS[k] - Moy_ar
MoyPlan_cycleAS[k] <- sum(Ys$Ys_cycleAS[,1])/Ys$n
ECPlan_cycleAS[k] <- MoyPlan_cycleAS[k] - Moy_cycle
MoyPlan_normST[k] <- sum(Ys$Ys_normST[,1])/Ys$n
ECPlan_normST[k] <- MoyPlan_normST[k] - Moy_norm
MoyPlan_arST[k] <- sum(Ys$Ys_arST[,1])/Ys$n
ECPlan_arST[k] <- MoyPlan_arST[k] - Moy_ar
MoyPlan_cycleST[k] <- sum(Ys$Ys_cycleST[,1])/Ys$n
ECPlan_cycleST[k] <- MoyPlan_cycleST[k] - Moy_cycle
MoyPlan_normNS[k] <- sum(Ys$Ys_normNS[,1])/Ys$n

```

```

ECPlan_normNS[k] <- MoyPlan_normNS[k] - Moy_norm
MoyPlan_arNS[k] <- sum(Ys$Ys_arNS[,1])/Ys$n
ECPlan_arNS[k] <- MoyPlan_arNS[k] - Moy_ar
MoyPlan_cycleNS[k] <- sum(Ys$Ys_cycleNS[,1])/Ys$n
ECPlan_cycleNS[k] <- MoyPlan_cycleNS[k] - Moy_cycle
}

PropCompl_AS <- PropCompl_AS/nrep
PropCompl_ST <- PropCompl_ST/nrep
PropCompl_NS <- PropCompl_NS/nrep

# Calcul des erreurs quadratiques moyennes (EQM)
# Par rapport au Modèle
EQMmod_normAS <- mean(ECmod_normAS)
EQMmod_normST <- mean(ECmod_normST)
EQMmod_normNS <- mean(ECmod_normNS)
EQMmod_arAS <- mean(ECmod_arAS)
EQMmod_arST <- mean(ECmod_arST)
EQMmod_arNS <- mean(ECmod_arNS)
EQMmod_cycleAS <- mean(ECmod_cycleAS)
EQMmod_cycleST <- mean(ECmod_cycleST)
EQMmod_cycleNS <- mean(ECmod_cycleNS)

# Par rapport au Plan de sondage
EQMPlan_normAS <- mean(ECPlan_normAS)
EQMPlan_normST <- mean(ECPlan_normST)
EQMPlan_normNS <- mean(ECPlan_normNS)
EQMPlan_arAS <- mean(ECPlan_arAS)
EQMPlan_arST <- mean(ECPlan_arST)
EQMPlan_arNS <- mean(ECPlan_arNS)
EQMPlan_cycleAS <- mean(ECPlan_cycleAS)
EQMPlan_cycleST <- mean(ECPlan_cycleST)
EQMPlan_cycleNS <- mean(ECPlan_cycleNS)

# Calcul du biais des estimateurs relativement à la moyenne (BaisR) en pourcentage
# Par rapport au Modèle
BiaisRmod_normAS <- 100*mean(ECmod_normAS)/Moy_norm
BiaisRmod_normST <- 100*mean(ECmod_normST)/Moy_norm
BiaisRmod_normNS <- 100*mean(ECmod_normNS)/Moy_norm
BiaisRmod_arAS <- 100*mean(ECmod_arAS)/Moy_ar
BiaisRmod_arST <- 100*mean(ECmod_arST)/Moy_ar
BiaisRmod_arNS <- 100*mean(ECmod_arNS)/Moy_ar
BiaisRmod_cycleAS <- 100*mean(ECmod_cycleAS)/Moy_cycle

```

```

BiaisRmod_cycleST <- 100*mean(ECmod_cycleST)/Moy_cycle
BiaisRmod_cycleNS <- 100*mean(ECmod_cycleNS)/Moy_cycle

# Par rapport au Plan de sondage
BiaisRPlan_normAS <- 100*mean(ECPlan_normAS)/Moy_norm
BiaisRPlan_normST <- 100*mean(ECPlan_normST)/Moy_norm
BiaisRPlan_normNS <- 100*mean(ECPlan_normNS)/Moy_norm
BiaisRPlan_arAS <- 100*mean(ECPlan_arAS)/Moy_ar
BiaisRPlan_arST <- 100*mean(ECPlan_arST)/Moy_ar
BiaisRPlan_arNS <- 100*mean(ECPlan_arNS)/Moy_ar
BiaisRPlan_cycleAS <- 100*mean(ECPlan_cycleAS)/Moy_cycle
BiaisRPlan_cycleST <- 100*mean(ECPlan_cycleST)/Moy_cycle
BiaisRPlan_cycleNS <- 100*mean(ECPlan_cycleNS)/Moy_cycle

# Calcul de la racine carrée des erreurs quadratiques moyennes relativement à la moyenne
(EQMR) en pourcentage
# Par rapport au Modèle
EQMRmod_normAS <- 100*sqrt(mean(ECmod_normAS^2))/Moy_norm
EQMRmod_normST <- 100*sqrt(mean(ECmod_normST^2))/Moy_norm
EQMRmod_normNS <- 100*sqrt(mean(ECmod_normNS^2))/Moy_norm
EQMRmod_arAS <- 100*sqrt(mean(ECmod_arAS^2))/Moy_ar
EQMRmod_arST <- 100*sqrt(mean(ECmod_arST^2))/Moy_ar
EQMRmod_arNS <- 100*sqrt(mean(ECmod_arNS^2))/Moy_ar
EQMRmod_cycleAS <- 100*sqrt(mean(ECmod_cycleAS^2))/Moy_cycle
EQMRmod_cycleST <- 100*sqrt(mean(ECmod_cycleST^2))/Moy_cycle
EQMRmod_cycleNS <- 100*sqrt(mean(ECmod_cycleNS^2))/Moy_cycle

# Par rapport au Plan de sondage
EQMRPlan_normAS <- 100*sqrt(mean(ECPlan_normAS^2))/Moy_norm
EQMRPlan_normST <- 100*sqrt(mean(ECPlan_normST^2))/Moy_norm
EQMRPlan_normNS <- 100*sqrt(mean(ECPlan_normNS^2))/Moy_norm
EQMRPlan_arAS <- 100*sqrt(mean(ECPlan_arAS^2))/Moy_ar
EQMRPlan_arST <- 100*sqrt(mean(ECPlan_arST^2))/Moy_ar
EQMRPlan_arNS <- 100*sqrt(mean(ECPlan_arNS^2))/Moy_ar
EQMRPlan_cycleAS <- 100*sqrt(mean(ECPlan_cycleAS^2))/Moy_cycle
EQMRPlan_cycleST <- 100*sqrt(mean(ECPlan_cycleST^2))/Moy_cycle
EQMRPlan_cycleNS <- 100*sqrt(mean(ECPlan_cycleNS^2))/Moy_cycle

obj <- list(MoyMod_normAS=MoyMod_normAS, MoyMod_arAS=MoyMod_arAS,
MoyMod_cycleAS=MoyMod_cycleAS, MoyMod_normST=MoyMod_normST,
MoyMod_arST=MoyMod_arST, MoyMod_cycleST=MoyMod_cycleST,
MoyMod_normNS=MoyMod_normNS, MoyMod_arNS=MoyMod_arNS,

```

```

MoyMod_cycleNS=MoyMod_cycleNS, MoyPlan_normAS=MoyPlan_normAS,
MoyPlan_arAS=MoyPlan_arAS, MoyPlan_cycleAS=MoyPlan_cycleAS,
MoyPlan_normST=MoyPlan_normST, MoyPlan_arST=MoyPlan_arST,
MoyPlan_cycleST=MoyPlan_cycleST, MoyPlan_normNS=MoyPlan_normNS,
MoyPlan_arNS=MoyPlan_arNS, MoyPlan_cycleNS=MoyPlan_cycleNS,
ECmod_normAS=ECmod_normAS, ECmod_arAS=ECmod_arAS,
ECmod_cycleAS=ECmod_cycleAS, ECmod_normST=ECmod_normST,
ECmod_arST=ECmod_arST, ECmod_cycleST=ECmod_cycleST,
ECmod_normNS=ECmod_normNS, ECmod_arNS=ECmod_arNS,
ECmod_cycleNS=ECmod_cycleNS, ECPlan_normAS=ECPlan_normAS,
ECPlan_arAS=ECPlan_arAS, ECPlan_cycleAS=ECPlan_cycleAS,
ECPlan_normST=ECPlan_normST, ECPlan_arST=ECPlan_arST,
ECPlan_cycleST=ECPlan_cycleST, ECPlan_normNS=ECPlan_normNS,
ECPlan_arNS=ECPlan_arNS, ECPlan_cycleNS=ECPlan_cycleNS,
BiaisRmod_normAS=BiaisRmod_normAS, BiaisRmod_arAS=BiaisRmod_arAS,
BiaisRmod_cycleAS=BiaisRmod_cycleAS, BiaisRmod_normST=BiaisRmod_normST,
BiaisRmod_arST=BiaisRmod_arST, BiaisRmod_cycleST=BiaisRmod_cycleST,
BiaisRmod_normNS=BiaisRmod_normNS, BiaisRmod_arNS=BiaisRmod_arNS,
BiaisRmod_cycleNS=BiaisRmod_cycleNS, BiaisRPlan_normAS=BiaisRPlan_normAS,
BiaisRPlan_arAS=BiaisRPlan_arAS, BiaisRPlan_cycleAS=BiaisRPlan_cycleAS,
BiaisRPlan_normST=BiaisRPlan_normST, BiaisRPlan_arST=BiaisRPlan_arST,
BiaisRPlan_cycleST=BiaisRPlan_cycleST, BiaisRPlan_normNS=BiaisRPlan_normNS,
BiaisRPlan_arNS=BiaisRPlan_arNS, BiaisRPlan_cycleNS=BiaisRPlan_cycleNS,
EQMRmod_normAS=EQMRmod_normAS, EQMRmod_arAS=EQMRmod_arAS,
EQMRmod_cycleAS=EQMRmod_cycleAS, EQMRmod_normST=EQMRmod_normST,
EQMRmod_arST=EQMRmod_arST, EQMRmod_cycleST=EQMRmod_cycleST,
EQMRmod_normNS=EQMRmod_normNS, EQMRmod_arNS=EQMRmod_arNS,
EQMRmod_cycleNS=EQMRmod_cycleNS, EQMRPlan_normAS=EQMRPlan_normAS,
EQMRPlan_arAS=EQMRPlan_arAS, EQMRPlan_cycleAS=EQMRPlan_cycleAS,
EQMRPlan_normST=EQMRPlan_normST, EQMRPlan_arST=EQMRPlan_arST,
EQMRPlan_cycleST=EQMRPlan_cycleST, EQMRPlan_normNS=EQMRPlan_normNS,
EQMRPlan_arNS=EQMRPlan_arNS, EQMRPlan_cycleNS=EQMRPlan_cycleNS,
Moy_norm=Moy_norm, Moy_ar=Moy_ar, Moy_cycle=Moy_cycle,
PropCompl_AS=PropCompl_AS, PropCompl_ST=PropCompl_ST,
PropCompl_NS=PropCompl_NS, nrep=nrep)
}

```

# Bibliographie

- Biemer P. and Stokes L. (1985). Optimal Design of Interviewer Variance Experiments in Complex Surveys, *Journal of the American Statistical Association*, 80, 158-166.
- Biemer P. and Stokes L. (1989). The Optimal Design of Quality Control Samples to Detect Interviewer Cheating, *Journal of Official Statistics*, 5, 23-39.
- Biemer P. and Stokes L. (1991). Approaches to The Modeling of Measurement Errors, in Biemer P., Groves R., Lyberg L., Mathiowetz N. and Sudman S. (eds), *Mesurement Errors in Surveys*. Wiley, New York pp. 487-516.
- Brewer K. R. W. et Sarndal C. E. (1983). Six Approaches to Enumerative Survey Sampling, in Madow W. G. and Olkin I. (eds) . *Incomplete Data in Sample Surveys*, vol. 3, Academic Press, pp.363-368.
- Gentle J. (1998). *Numerical Linear Algebra for Applications in Statistics*. Springer-Verlag, New York.
- Hasen M. H., Madow W. G. et Tepping, B. J. (1983). An Evaluation of Model-Dependant and Probability Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, 78, 777-793.
- Huber P. J. (1981). *Robust Statistics*. Wiley, New York.
- Judith T. L. et William D. K. (1992). *Nonsampling Error in Surveys*. Wiley, New York.
- Montgomery D. C. (2001). *Design and Analysis of Experiments*. Wiley, New York.
- Royall R. M. et Herson J. (1973a). Robust Estimation in Finite Populations I. *Journal of the American Statistical Association*, 68, 880-889.
- Royall R. M. (1976b). The Linear Least Square Prediction Approach to Two-Sample Sampling, *Journal of the American Statistical Association*, 71, 657-664.
- Searle S. R. (1971). *Linear Models*. Wiley, New York.
- Sharon L. L. (1999). *Sampling : Design and Analysis*. Duxbury Press, Californie.
- Valliant R., Dorfman A. H., et Royall M. R. (2000). *Finite Population Sampling and Inference A Prediction Approach.*, Wiley, New York.