REVUE DE STATISTIQUE APPLIQUÉE

D. CAUSEUR

Plan d'échantillonnage en plusieurs phases pour la réduction des coûts expérimentaux en régression linéaire

Revue de statistique appliquée, tome 46, n° 4 (1998), p. 59-73

<http://www.numdam.org/item?id=RSA_1998__46_4_59_0>

Rev. Statistique Appliquée, 1998, **XLVI** (4), 59-73

PLAN D'ÉCHANTILLONNAGE EN PLUSIEURS PHASES POUR LA RÉDUCTION DES COÛTS EXPÉRIMENTAUX EN RÉGRESSION LINÉAIRE

D. Causeur

Institut National de la Recherche Agronomique - Laboratoire de Biométrie
Ecole Nationale Supérieure Agronomique de Rennes -
Laboratoire de Mathématiques Appliquées, 65 Rue de St-Brieuc 35042 Rennes, France

RÉSUMÉ

L'utilisation d'une variable auxiliaire dans un protocole d'échantillonnage à deux phases est suggérée par Cochran (1963) dans le but de réduire les coûts expérimentaux pour l'estimation de la moyenne d'une variable. La généralisation de cette méthode à l'estimation dans des modèles de régression linéaire est due à Conniffe et Moran (1972). Une nouvelle extension à plusieurs variables auxiliaires est proposée par Causeur et Dhorne (1998) sur la base des résultats obtenus par Conniffe (1985) dans le cadre des équations simultanées en économétrie. L'objectif de cet article est de généraliser la procédure d'optimisation du plan d'échantillonnage à un protocole en plusieurs phases. Les résultats obtenus sont illustrés par la planification d'une expérience préalable à l'élaboration d'une équation de prédiction du taux de muscle dans des carcasses de porcs.

Mots-clés : *Double-échantillonnage, Echantillonnage en plusieurs phases, Régression linéaire, Coût expérimental, Classement de carcasses de porcs.*

ABSTRACT

The use of an auxiliary covariate in a double-sampling scheme was suggested by Cochran (1963) in order to reduce experimental costs for the estimation of the expectation of a variable. The generalization of this method to the case of linear regression is due to Conniffe and Moran (1972). A further extension to the multivariate case was proposed by Causeur and Dhorne (1998) on the basis of the results obtained by Conniffe (1985) in the context of simultaneous equations in econometrics. The major aim of this paper is to widen the optimization procedure to a multiphase sampling design. Theoretical results are illustrated by the prediction of the lean meat percentage in pigs carcasses.

Keywords : *Two-phase sampling, Multiphase sampling, Linear regression, Experimental cost, Pig carcasses grading.*

1. Introduction

De nombreux développements méthodologiques visant à réduire les coûts expérimentaux s'appuient sur un protocole d'échantillonnage à plusieurs phases. A l'origine de ces méthodes, Cochran (1963) suggère de substituer à des mesures trop coûteuses d'une variable dont on cherche à estimer la moyenne dans une population à taille finie, des mesures de coût plus faible d'une variable auxiliaire fortement corrélée avec la variable d'intérêt. Le schéma d'échantillonnage qui en résulte naturellement fait intervenir un échantillon sur lequel seule la variable auxiliaire est mesurée et un sous-échantillon sur lequel la variable d'intérêt est mesurée conjointement à la variable auxiliaire. Une première extension de cette procédure à l'estimation des paramètres d'un modèle de régression est due à Conniffe et Moran (1972). Les propriétés de cette méthode, appelée double-régression, dans un cadre asymptotique, font l'objet d'une étude de Engel et Walstra (1991). La loi exacte des estimateurs est donnée par Causeur (1998) dans le but de définir des stratégies d'optimisation du plan d'échantillonnage valides pour des petites tailles d'échantillon.

Les extensions récentes de cette méthode portent sur l'introduction dans le modèle de plusieurs variables auxiliaires. Une première approche, due à Cook *et al.* (1983), vise à sélectionner dans un ensemble de variables auxiliaires celle qui réalise le meilleur compromis entre un coût d'observation faible et une corrélation élevée avec la variable d'intérêt. Une autre approche, décrite par Conniffe (1985), consiste à introduire simultanément plusieurs variables auxiliaires dans un schéma de double-échantillonnage. Causeur et Dhorne (1998) proposent une étude dans un contexte non-asymptotique des propriétés des estimateurs de la double-régression multivariable et compare les approches de Cook *et al.* (1983) et Conniffe (1985) dans le contexte de la réduction des coûts expérimentaux.

Ces méthodes statistiques sont notamment utilisées pour la planification des expériences intervenant dans le classement des carcasses de porcs. L'Union européenne définit en effet comme seul critère objectif de classement des carcasses de porcs leur taux de muscle. N'étant pas observable en conditions industrielles, ce taux de muscle est prédit à partir de mesures rapides d'épaisseurs de gras et de muscle. Les expériences préalables à l'élaboration des formules de prédiction sont fortement contraintes par le coût élevé d'observation du taux de muscle, à savoir le coût de la dissection complète d'une carcasse. Des variables construites à partir de dissections partielles et donc moins coûteuses sont alors utilisées comme variables auxiliaires dans un protocole de double-régression.

L'objectif de cet article est de présenter une généralisation de la double-régression multivariable à un schéma d'échantillonnage en plusieurs phases. Cette extension permet en particulier la prise en compte des coûts d'observation de chaque sous-ensemble de variables auxiliaires dans la définition d'un plan d'échantillonnage minimisant le coût global de l'expérience. Dans un premier temps, on présente le modèle et les liens fonctionnels intervenant entre les paramètres. On donne ensuite la forme explicite des estimateurs du maximum de vraisemblance des paramètres d'intérêt dans le cas gaussien. On propose enfin un critère d'efficacité relative permettant de comparer la méthode basée sur un échantillonnage à plusieurs phases et la méthode des moindres carrés ordinaires. L'utilisation de ce critère d'efficacité relative est illustrée par la prédiction du taux de muscle de carcasses de porcs.

2. Modélisation

2.1. Modèle d'intérêt et modèles auxiliaires

Soit Y la variable d'intérêt et X le vecteur ligne à p composantes des variables prédictrices, on suppose :

$$E(Y) = X\beta,\ V(Y) = \sigma^2, \tag{1}$$

où $\beta \in \mathbf{R}^p$ et $\sigma^2 \in \mathbf{R}^+$. Dans la suite, β et σ^2 sont appelés paramètres d'intérêt du modèle.

Soit $Z = (\ Z_1 \quad Z_2 \quad \ldots \quad Z_q\ )$ le vecteur ligne à q composantes des variables auxiliaires, on suppose d'une part :

$$E(Y|Z) = X\beta_{y|z} + Z\gamma_{y|z},\ V(Y|Z) = \sigma^2_{y|z}, \tag{2}$$

où $\beta_{y|z} \in \mathbf{R}^p$, $\gamma_{y|z} \in \mathbf{R}^q$, et $\sigma^2_{y|z} \in \mathbf{R}^+$, et d'autre part, pour tout $i \in \{2, \ldots, q\}$:

$$E(Z_i|Z^{(i-1)}) = X\beta_{i|i-1} + Z^{(i-1)}\gamma_{i|i-1},\ V(Z_i|Z^{(i-1)}) = \sigma^2_{i|i-1}, \tag{3}$$

où $Z^{(i-1)}$ est le vecteur ligne constitué des $i-1$ premières composantes de Z, $\beta_{i|i-1} \in \mathbf{R}^p$, $\gamma_{i|i-1} \in \mathbf{R}^{i-1}$ et $\sigma^2_{i|i-1} \in \mathbf{R}^+$. Par extension, pour $i = 1$, on suppose aussi que :

$$E(Z_1) = X\beta_{1|0},\ V(Z_1) = \sigma^2_{1|0}, \tag{4}$$

où $\beta_{1|0} \in \mathbf{R}^p$ et $\sigma^2_{1|0} \in \mathbf{R}^+$.

Les modèles décrits par les relations (2), (3) et (4) sont appelés dans la suite modèles auxiliaires.

2.2. Liens entre paramètres d'intérêt et paramètres des modèles auxiliaires

Les relations (3) et (4) peuvent aussi se mettre sous la forme suivante :

$$E\left[Z_i \mid Z^{(i-1)}\right] = X\beta_{i|i-1} + Z\gamma^*_{i|i-1},\ i = 1, \ldots, q,$$

où $\gamma^*_{i|i-1} \in \mathbf{R}^q$ est obtenu en complétant le vecteur $\gamma_{i|i-1} \in \mathbf{R}^{i-1}$ par des 0. Par souci d'homogénéité des notations, $\gamma^*_{1|0}$ est donc le vecteur nul.

On en déduit que :

$$E[Z_i] = X\beta_{i|i-1} + E[Z]\gamma^*_{i|i-1},\ i = 1, \ldots, q.$$

Sous une forme matricielle, le système précédent devient :

$$E[Z] = X\beta_z + E[Z]\gamma_z, \tag{5}$$

où $\beta_z = [\beta_{1|0} \quad \beta_{2|1} \quad \cdots \quad \beta_{q|q-1}] \in \mathcal{M}_{p,q}(\mathbf{R})$ et $\gamma_z = \begin{bmatrix} \gamma^*_{1|0} & \gamma^*_{2|1} & \cdots & \gamma^*_{q|q-1} \end{bmatrix} \in \mathcal{M}_{q,q}(\mathbf{R})$ est une matrice triangulaire supérieure, à diagonale nulle.

On déduit de (5) que :

$$E[Z] = X\beta_z [I - \gamma_z]^{-1}, \tag{6}$$

Cette dernière relation conduit à une nouvelle expression de l'espérance de Y :

$$\begin{aligned} E[Y] &= E\{E[Y|Z]\}, \\ &= X\beta_{y|z} + E[Z]\gamma_{y|z}, \\ &= X\left(\beta_{y|z} + \beta_z [I - \gamma_z]^{-1} \gamma_{y|z}\right). \end{aligned}$$

Par identification de la relation précédente et de l'expression (1), on déduit le lien suivant entre paramètres d'intérêt et paramètres des modèles auxiliaires :

$$\beta = \beta_{y|z} + \beta_z [I - \gamma_z]^{-1} \gamma_{y|z}. \tag{7}$$

Dans la suite, pour tout $i \in \{1, 2, \ldots, q\}$, $\Sigma_i \in \mathcal{M}_{i,i}(\mathbf{R})$ désigne la variance du vecteur $Z^{(i)}$. D'après la formule de la variance totale et les relations (3) et (4) :

$$\begin{aligned} V\left[Z^{(i)}\right] &= E\left\{V\left[Z^{(i)}|Z^{(i-1)}\right]\right\} + V\left\{E\left[Z^{(i)}|Z^{(i-1)}\right]\right\}, \\ &= \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{i|i-1} \end{pmatrix} + V\left\{\begin{pmatrix} Z^{(i-1)} \\ Z^{(i-1)}\gamma_{i|i-1} \end{pmatrix}\right\}, \\ &= \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2_{i|i-1} \end{pmatrix} + \begin{pmatrix} \Sigma_{i-1} & \Sigma_{i-1}\gamma_{i|i-1} \\ \gamma'_{i|i-1}\Sigma_{i-1} & \gamma'_{i|i-1}\Sigma_{i-1}\gamma_{i|i-1} \end{pmatrix}. \end{aligned}$$

On en déduit la relation de récurrence suivante entre deux termes successifs de la suite $(\Sigma_i)_{i \geqslant 1}$:

$$\Sigma_i = \begin{pmatrix} \Sigma_{i-1} & \Sigma_{i-1}\gamma_{i|i-1} \\ \gamma'_{i|i-1}\Sigma_{i-1} & \sigma^2_{i|i-1} + \gamma'_{i|i-1}\Sigma_{i-1}\gamma_{i|i-1} \end{pmatrix}. \tag{8}$$

Or, d'après la formule de la variance totale,

$$\begin{aligned} V[Y] &= E\{V[Y|Z]\} + V\{E[Y|Z]\}, \\ &= \sigma^2_{y|z} + V\{Z\gamma_{y|z}\}, \\ &= \sigma^2_{y|z} + \gamma'_{y|z} V(Z) \gamma_{y|z}. \end{aligned}$$

La relation précédente conduit à une nouvelle expression de σ^2 :

$$\sigma^2 = \sigma^2_{y|z} + \gamma'_{y|z}\Sigma_q\gamma_{y|z}. \tag{9}$$

Comme d'autre part, d'après la relation de récurrence (8), Σ_q peut s'exprimer uniquement en fonction des paramètres des modèles (3) et (4), l'expression précédente définit le lien entre le paramètre d'intérêt σ^2 et les paramètres des modèles auxiliaires.

Dans le cas où $q = 1$, Z désigne l'unique variable auxiliaire, $\gamma_{y|z} \in \mathbf{R}$, $\gamma_z = 0$, $\beta_z \in \mathbf{R}^p$ et $\Sigma_q = \sigma^2_z \in \mathbf{R}^+$. Les relations (7) et (9) deviennent alors :

$$\beta = \beta_{y|z} + \gamma_{y|z}\beta_z,$$
$$\sigma^2 = \sigma^2_{y|z} + \gamma^2_{y|z}\sigma^2_z.$$

2.3. Plan d'échantillonnage

Le plan d'échantillonage utilisé pour estimer les paramètres d'intérêt fait intervenir $q + 1$ échantillons selon un protocole monotone :

- un échantillon de taille n_1 pour lequel Z_1 et les variables prédictrices sont conjointement observables,
- un sous-échantillon de taille n_2, avec $n_2 \leqslant n_1$, constitué de toutes les unités statistiques du premier échantillon pour lesquelles Z_2 est observable,
- ...
- un sous-échantillon de taille n_{q+1}, avec $n_{q+1} \leqslant n_q$, constitué de toutes les unités statistiques du sous-échantillon précédent pour lesquelles toutes les variables du modèle sont observables; en particulier Y n'est observable que pour ce dernier sous-échantillon.

La monotonie de ce plan d'échantillonnage traduit en pratique une hiérarchie entre les variables du modèle : plus le coût d'observation d'une variable est élevé, plus la taille de l'échantillon sur lequel cette variable est observée est faible.

Comme il a été mentionné par Causeur et Dhorne (1998), la vraisemblance du modèle global peut être calculée dans le cas où les variables prédictrices sont l'objet d'un contrôle expérimental. Dans le cas d'une observation séquentielle des variables Z_i, $i = 1, \ldots, q$, une procédure de sélection des individus observés est donc envisageable, consistant à choisir au vu de la réalisation conjointe des $i - 1$ premières variables sur l'échantillon de taille n_{i-1}, les unités pour lesquelles Z_i est observée.

Dans la suite, $Y^{(n_{q+1})}$ désigne le vecteur des observations de Y sur l'échantillon de taille n_{q+1}, $Z_j^{(n_i)}$, $i = 1, \ldots, q$, $j = 1, \ldots, i$, le vecteur des observations de Z_j sur l'échantillon de taille n_i et $X^{(n_i)}$ la matrice des observations de X sur l'échantillon de taille n_i, $i = 1, \ldots, q$.

3. Estimation

On suppose ici que le vecteur $(Y, Z_1, Z_2, \ldots, Z_q)'$ est distribué selon une loi normale et on propose d'estimer les paramètres d'intérêt β et σ^2 par la méthode du maximum de vraisemblance. Le calcul suivant des estimateurs du maximum de vraisemblance des paramètres d'intérêt généralise le calcul de Engel et Walstra (1991) dans un contexte univariable ($q = 1$).

D'après les relations (7) et (9), on peut considérer comme ensemble des paramètres intrinsèques du modèle l'ensemble θ suivant des paramètres intervenant dans les modèles auxiliaires (2), (3) et (4) :

$$\theta = \theta_{y|z} \cup \left\{ \bigcup_{i=1}^{q} \theta_{i|i-1} \right\},$$

où $\theta_{y|z} = \left\{ \beta_{y|z}, \gamma_{y|z}, \sigma^2_{y|z} \right\}$, $\theta_{i|i-1} = \left\{ \beta_{i|i-1}, \gamma_{i|i-1}, \sigma^2_{i|i-1} \right\}$, pour $i = 2, \ldots, q$ et $\theta_{1|0} = \left\{ \beta_{1|0}, \sigma^2_{1|0} \right\}$.

La log-vraisemblance $L\left(y^{(n_{q+1})}, z_i^{(n_i)},\ i = 1, \ldots, q; \theta \right)$ est la somme des $q + 1$ log-vraisemblances $L_{y|z}\left(y^{(n_{q+1})}; z_i^{(n_i)},\ i = 1, \ldots, q; \theta \right)$, $L_{i|i-1}\left(z_i^{(n_i)}; z_{i-1}^{(n_{i-1})}, \ldots, z_1^{(n_1)}; \theta \right)$, $i = 2, \ldots, q$ et $L_{1|0}\left(z_1^{(n_1)}; \theta \right)$ des modèles auxiliaires (2), (3) et (4). L'estimateur $\hat{\theta}$ du maximum de vraisemblance de θ vérifie donc la relation suivante :

$$\begin{aligned} \hat{\theta} = &\arg\max_{\theta} \left\{ L_{y|z}\left(y^{(n_{q+1})}; z_i^{(n_i)},\ i = 1, \ldots, q, \theta_{y|z} \right) \right. \\ &\left. + \sum_{i=1}^{q} L_{i|i-1}\left(z_i^{(n_i)}; z_{i-1}^{(n_{i-1})}, \ldots, z_1^{(n_1)}, \theta_{i|i-1} \right) \right\}. \end{aligned}$$

Cette dernière expression permet de séparer la maximisation globale en θ en maximisations des log-vraisemblances associées aux modèles auxiliaires (2), (3) et (4) :

$$\hat{\theta} = \hat{\theta}_{y|z} \cup \left\{ \bigcup_{i=1}^{q} \hat{\theta}_{i|i-1} \right\},$$

où

$$\hat{\theta}_{y|z} = \arg\max_{\theta_{y|z}} L_{y|z}\left(y^{(n_{q+1})}; z_i^{(n_i)},\ i = 1, \ldots, q; \theta_{y|z} \right)$$

et

$$\hat{\theta}_{i|i-1} = \arg\max_{\theta_{i|i-1}} L_{i|i-1}\left(z_i^{(n_i)}; z_{i-1}^{(n_{i-1})}, \ldots, z_1^{(n_1)}; \theta_{i|i-1} \right).$$

Les modèles auxiliaires étant supposés gaussiens, $\hat{\theta}_{y|z}$ et $\hat{\theta}_{i|i-1}$, $i = 1, 2, \ldots, q$, sont donc les estimateurs des moindres carrés ordinaires des paramètres des modèles (2), (3) et (4) calculés respectivement sur les échantillons de taille n_{q+1} et n_i, $i = 1, 2, \ldots, q$.

On déduit de la propriété d'invariance fonctionnelle de la méthode du maximum de vraisemblance et des relations (7) et (9) l'expression des estimateurs des paramètres d'intérêt :

$$\hat{\beta} = \hat{\beta}_{y|z} + \hat{\beta}_z \left[I - \hat{\gamma}_z\right]^{-1} \hat{\gamma}_{y|z},$$
$$\hat{\sigma}^2 = \hat{\sigma}^2_{y|z} + \hat{\gamma}'_{y|z} \hat{\Sigma}_q \hat{\gamma}_{y|z},$$

où $\hat{\beta}_z = \left[\hat{\beta}_{1|0} \quad \hat{\beta}_{2|1} \quad \ldots \quad \hat{\beta}_{q|q-1}\right] \in \mathcal{M}_{p,q}(\mathbf{R})$, $\hat{\gamma}_z = \left[\hat{\gamma}^*_{1|0} \quad \hat{\gamma}^*_{2|1} \quad \cdots \quad \hat{\gamma}^*_{q|q-1}\right] \in \mathcal{M}_{q,q}(\mathbf{R})$, $\hat{\gamma}^*_{i|i-1} \in \mathbf{R}^q$ est obtenu en complétant le vecteur $\hat{\gamma}_{i|i-1} \in \mathbf{R}^{i-1}$ par des 0, $\hat{\Sigma}_q$ est le $q^{\text{ème}}$ terme de la suite $(\hat{\Sigma}_i)_{i \geqslant 1}$ définie par récurrence de la façon suivante :

$$\hat{\Sigma}_1 = \hat{\sigma}^2_{1|0},$$
$$\hat{\Sigma}_i = \begin{pmatrix} \hat{\Sigma}_{i-1} & \hat{\Sigma}_{i-1}\hat{\gamma}_{i|i-1} \\ \hat{\gamma}'_{i|i-1}\hat{\Sigma}_{i-1} & \hat{\sigma}^2_{i|i-1} + \hat{\gamma}'_{i|i-1}\hat{\Sigma}_{i-1}\hat{\gamma}_{i|i-1} \end{pmatrix}.$$

L'étude des propriétés de ces estimateurs permet dans la suite de définir un critère de comparaison des efficacités de la méthode construite à partir d'un plan d'échantillonnage à plusieurs phases et de la méthode des moindres carrés ordinaires.

4. Efficacité relative

Dans le contexte de l'estimation d'une moyenne dans une population à taille finie, Cochran (1977) propose de mesurer les efficacités de l'estimateur basé sur un échantillonnage à deux phases et de la moyenne arithmétique par le rapport des variances de ces estimateurs. Une extension de ce critère repose donc naturellement sur le calcul de la variance de $\hat{\beta}$. Par souci de clarté, la démonstration du non-biais de $\hat{\beta}$ et le calcul de sa variance dans des conditions non-asymptotiques sont reportés en annexe.

Soit $S_{n_i} = \left[\left\{X^{(n_i)}\right\}' X^{(n_i)}\right]^{-1}$, $i = 1, \ldots, q+1$, soit $\sigma^2_{y|i} = V(Y|Z^{(i)})$, $i = 1, \ldots, q$, et $\sigma^2_{y|0} = \sigma^2$, alors

$$V\left[\hat{\beta}\right] = \sum_{j=1}^{q+1} \alpha_j S_{n_j},$$

où $\alpha = (\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_{q+1})' \in \mathbf{R}^{q+1}$, terme d'indice 0 de la suite $\alpha^i \in \mathbf{R}^{q-i+1}$, $i = 0, 1, \dots, q$, définie par récurrence décroissante de la façon suivante :

$$\begin{aligned} \alpha^q &= \sigma^2_{y|z}, \\ \alpha^i &= u_i + M_i \alpha^{i+1},\ i = 0, 1, \dots, q-1, \end{aligned}$$

avec $u_i = \left(\sigma^2_{y|i} - \sigma^2_{y|i+1} \quad 0 \quad \dots \quad 0\right)' \in \mathbf{R}^{q-i+1}$, $i = 0, 1, \dots, q-1$, et $M_i \in \mathcal{M}_{q-i+1,q-i}(\mathbf{R})$ est définie de la façon suivante :

$$M_i = \begin{pmatrix} -\dfrac{1}{n_{i+2}-p-i-3} & -\dfrac{1}{n_{i+3}-p-i-3} & \dots & -\dfrac{1}{n_{q+1}-p-i-3} \\ \dfrac{n_{i+2}-p-i-2}{n_{i+2}-p-i-3} & 0 & \dots & 0 \\ 0 & \dfrac{n_{i+3}-p-i-2}{n_{i+3}-p-i-3} & \dots & 0 \\ 0 & 0 & \dots & \dfrac{n_{q+1}-p-i-2}{n_{q+1}-p-i-3} \end{pmatrix},$$

$i = 0, 1, \dots, q-1$.

Dans le cas particulier où $q = 1$, on en déduit que :

$$V\left[\hat{\beta}\right] = \alpha_1 S_{n_1} + \alpha_2 S_{n_2},$$

où $\alpha = (\alpha_1 \quad \alpha_2)'$ vérifie la relation suivante :

$$\begin{aligned} \alpha &= u_0 + M_0 u_1, \\ &= \begin{pmatrix} \sigma^2 - \sigma^2_{y|z} \\ 0 \end{pmatrix} + \begin{pmatrix} -\dfrac{1}{n_2-p-3} \\ \dfrac{n_2-p-2}{n_2-p-3} \end{pmatrix} \sigma^2_{y|z}, \\ &= \begin{pmatrix} \sigma^2 - \dfrac{n_2-p-2}{n_2-p-3}\sigma^2_{y|z} \\ \dfrac{n_2-p-2}{n_2-p-3}\sigma^2_{y|z} \end{pmatrix}. \end{aligned}$$

On retrouve donc l'expression de la variance de $\hat{\beta}$ obtenue par Conniffe (1985) dans le cas d'un échantillonnage à deux phases :

$$V\left[\hat{\beta}\right] = \sigma^2 S_{n_1} + \sigma^2_{y|z} \frac{n_2-p-2}{n_2-p-3}\left(S_{n_2} - S_{n_1}\right).$$

Le critère d'efficacité relative exact proposé par Causeur et Dhorne (1998) dans le contexte d'un échantillonnage à deux phases est le rapport maximal des variances d'une combinaison linéaire arbitraire des estimateurs des coefficients de la

régression par la méthode s'appuyant sur l'échantillonnage à deux phases phases et par la méthode des moindres carrés ordinaires. La généralisation de ce critère à un échantillonnage en plusieurs phases s'exprime sous la forme suivante :

$$ERE(\beta) = \sup_{\lambda \in R^p} \frac{V\left(\lambda' \hat{\beta}\right)}{V\left(\lambda' \hat{\beta}_{MCO}\right)},$$

$$= \sup_{\lambda \in R^p} \frac{\sum_{j=1}^{q+1} \alpha_j \lambda' S_{n_j} \lambda}{\sigma^2 \lambda' S_m \lambda}, \tag{10}$$

où $\hat{\beta}$ est calculé à partir des échantillons de taille n_i, $i = 1, \ldots, q+1$ et $\hat{\beta}_{MCO}$ est l'estimateur par la méthode des moindres carrés ordinaires sur un échantillon de taille m. Le critère (10) peut aussi s'exprimer comme la plus grande valeur propre de la matrice :

$$\sum_{j=1}^{q+1} \frac{\alpha_j}{\sigma^2} S_m^{-1} S_{n_j}. \tag{11}$$

Le calcul effectif de ce critère suppose connues les valeurs prises par les variables prédictrices sur les échantillons de taille n_1 et m. Il est par conséquent utilisé pour mesurer l'efficacité relative après l'expérience et non pour planifier cette expérience. On peut cependant proposer un nouveau critère indépendant des valeurs prises par les variables prédictrices en remarquant que les matrices $(1/m)S_m^{-1}$ et $(1/n_j)S_{n_j}^{-1}$ sont deux estimateurs d'une même quantité. On en déduit une approximation du critère (10) valide pour de grandes tailles d'échantillon :

$$ERA(\beta) = \frac{m}{\sigma^2} \sum_{j=1}^{q+1} \frac{\alpha_j}{n_j}. \tag{12}$$

Le critère précédent ne dépend que des tailles d'échantillon et des carrés des coefficients de corrélation partielle multiple $\rho_{y,i}^2$, $i = 1, \ldots, q$, entre Y et $(Z_1, Z_2, \ldots, Z_i)$ sachant X, définis par :

$$\rho_{y,i}^2 = 1 - \frac{\sigma_{y|i}^2}{\sigma^2}.$$

5. Exemple : prédiction du taux de muscle de carcasses de porcs

Le critère (12) est utilisé dans la suite pour la planification d'expériences visant à établir une formule de prédiction du taux de muscle de carcasses de porcs.

5.1. Contraintes expérimentales

Le classement des carcasses de porcs dans l'ensemble des pays de l'Union Européenne est basé sur le taux de muscle de ces carcasses. Cependant, la mesure de cette quantité nécessite une dissection totale. Les coûts élevés et le caractère destructif de cette mesure ont deux conséquences directes :

- en conditions industrielles, le taux de muscle n'est pas observé mais prédit à partir de mesures rapides. En France, les deux variables prédictrices sont des épaisseurs de gras et de muscle mesurées le plus souvent par un appareil de mesure de réflectance tissulaire.
- l'expérience visant à établir l'équation de prédiction est aussi contrainte par le coût, en terme de temps de dissection, de la mesure du taux de muscle. Ces contraintes orientent le choix d'une méthodologie statistique vers des techniques d'estimation non-standard visant à réduire les coûts expérimentaux par l'introduction dans le modèle de variables obtenues par des dissections partielles. Dans l'optique de l'harmonisation européenne des méthodes de prédiction du taux de muscle, l'Union Européenne impose toutefois un niveau minimal d'efficacité de la méthode statistique utilisée correspondant à la précision moyenne de la méthode des moindres carrés ordinaires sur un échantillon de 120 carcasses.

Les développements suivants ont pour objectif de détailler la phase de planification du prochain essai expérimental devant avoir lieu en France sur la base de l'estimation, lors d'un essai préalable, des paramètres et des coûts intervenant dans la procédure d'optimisation du plan d'échantillonnage.

5.2. Planification de l'échantillonnage

Dans la suite, la variable d'intérêt, à savoir le taux de muscle est noté $TMUS$, et les variables prédictrices $EPGRAS$ et $EPMUS$. Le coût d'observation du taux de muscle est de 380 minutes et le coût d'observation des épaisseurs de gras et de muscle est considéré comme négligeable.

Dans un premier temps, un ensemble de variables auxiliaires est choisi sur la base du coût de leur mesure, en terme de temps de dissection, et de leur corrélation avec la variable d'intérêt. Cet ensemble est constitué de 7 variables : une variable D obtenue après découpe de la carcasse par différence entre le poids des pièces à dominante grasse et le poids des pièces à dominante maigre, le taux de muscle P dans la poitrine, le taux de muscle E dans l'épaule, le taux de muscle J dans le jambon, le taux de muscle Lc dans la partie costale de la longe, le taux de muscle Ld dans la partie dorsale de la longe, le taux de muscle F dans le filet. Les coûts de mesure des variables auxiliaires et le carré des coefficients de corrélation partielle entre $TMUS$ et chacune de ces variables connaissant $EPGRAS$ et $EPMUS$ figurent dans le tableau 1.

Les coûts de mesure d'un sous-ensemble de variables auxiliaires ne se déduisent pas par additivité des coûts unitaires de ses composantes, le travail nécessaire à la mesure de différentes variables auxiliaires pouvant être commun. Afin de visualiser l'effet de l'augmentation de la taille des sous-ensembles de variables auxiliaires, on a représenté sur les figures 1 et 2 respectivement les carrés de coefficients

TABLEAU 1

Coût unitaire et carré du coefficient de corrélation partielle entre $TMUS$ et chaque variable auxiliaire connaissant $(EPGRAS, EPMUS)$

Sous-ensemble	D	P	J	E	Lc	Ld	F
$\rho^2_{y,z}$	0.39	0.00	0.55	0.25	0.23	0.02	0.21
Coût unitaire	20	40	35	35	30	30	5

de corrélation multiple partielle entre $TMUS$ et un sous-ensemble de variables auxiliaires connaissant $EPGRAS$ et $EPMUS$ et les coûts unitaires maximaux et minimaux par taille de sous-ensemble de variables auxiliaires.

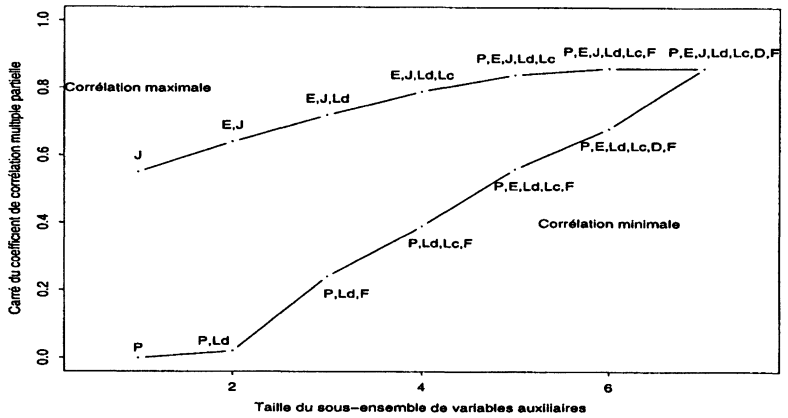

FIGURE 1

Carrés des coefficients de corrélation multiple partielle entre $TMUS$ et un sous-ensemble de variables auxiliaires, connaissant $EPGRAS$ et $EPMUS$, maximaux et minimaux pour une taille de sous-ensemble fixée.

Pour un sous-ensemble de q variables auxiliaires donné, on recense dans un premier temps toutes les combinaisons de tailles d'échantillons garantissant une méthodologie aussi précise que la méthode des moindres carré sur un échantillon de taille $m = 120$. En d'autres termes, ces combinaisons $(n_{q+1}, n_q, \ldots, n_1)$ vérifient $ERA(\beta) = 1$, où $ERA(\beta)$ est le critère d'efficacité relative défini par l'expression (12).

Formellement, n_{q+1} varie entre $p + q + 2$ et m. Pour chaque valeur de n_{q+1}, on fait alors varier n_q sous les contraintes suivantes : $n_q \geqslant n_{q+1}$ et $ERA(\beta) \leqslant 1$ où $ERA(\beta)$ est le critère d'efficacité relative défini avec le plan d'échantillonnage à $q+1$ phases $(n_{q+1}, n_q, \ldots, n_q)$. On définit de la même manière les plages de variation de $n_{q-1}, \ldots, n_1$.

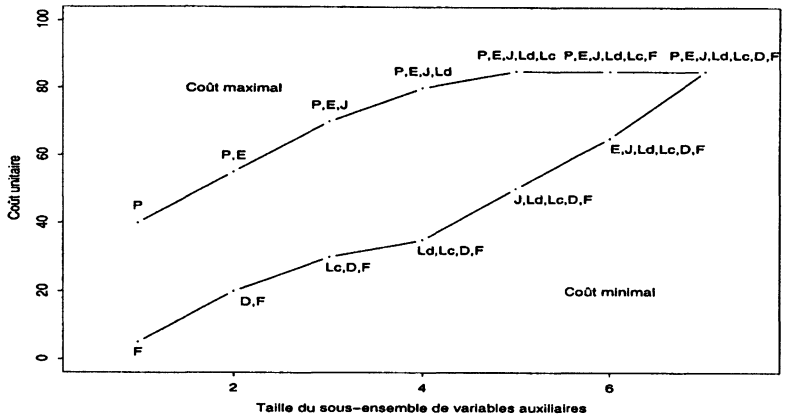

FIGURE 2

Coûts unitaires maximaux et minimaux de l'observation d'un sous-ensemble de variables auxiliaires par taille de sous-ensemble.

On choisit dans un second temps, parmi les plans d'échantillonnage retenus par l'algorithme précédent, celui qui est associé au coût expérimental le plus faible.

Le tableau 2 donne les plans d'échantillonnage optimaux et les coûts expérimentaux associés pour les meilleurs sous-ensembles de même taille. Les réductions de coût calculées dans ce tableau se rapportent au coût de la mise en œuvre de la méthode des moindres carrés sur un échantillon de taille 120, soit 120*380 = 45600 minutes de dissection.

TABLEAU 2

Plan d'échantillonnage optimal. Sous-ensembles optimaux : J, (J,F), (J,Ld,F), (E,J,Ld,Lc), (E,J,Ld,Lc,D), (E,J,Ld,Lc,D,F), (P,E,J,Ld,Lc,D,F). Pour une taille q fixée, on donne les effectifs dans l'ordre $(n_{q+1}, n_q, \ldots, n_1)$

Taille q	*Plan optimal*	*Coût*	*Réduction (%)*
0	120	45600	0
1	(76,245)	34795	23.70
2	(69,216,398)	32275	29.22
3	(64,195,195,377)	31125	31.74
4	(53,134,209,209,209)	29155	36.06
5	(48,153,178,178,178,241)	27575	39.53
6	(47,141,157,157,157,217,337)	26570	41.73
7	(46,61,140,155,155,155,215,333)	26430	42.04

6. Conclusion

La procédure de planification de l'échantillonnage décrite précédemment est inversible; on peut en effet choisir de minimiser la précision de l'estimation à ressources expérimentales fixées. D'autre part, quoique présentée dans le cadre de la régression linéaire, les résultats précédents généralisent, dans le cas où $p = 0$, le critère de Cochran (1977) pour l'estimation de la moyenne d'une variable d'intérêt dans une population de taille finie.

Dans le domaine du classement de carcasses de porcs, plusieurs pays membres de l'Union européenne utilisent l'information auxiliaire apportée par des dissections partielles dans un schéma de double-échantillonnage. L'exemple précédent montre l'intérêt d'affiner la procédure d'optimisation du plan d'échantillonnage en tenant compte des différences de coût entre les différentes variables auxiliaires.

Références

CAUSEUR D. et DHORNE T., (1998), Finite-sample properties of a multivariate extension of double-regression, *Biometrics*, A paraître.

CAUSEUR D, (1998), Exact distribution of the regression estimator in double-sampling, *Statistics*, A paraître.

COCHRAN W.G., (1963), Sampling Techniques, Second edition, *Wiley, New-York.*

COCHRAN W.G., (1977), Sampling Techniques, Third edition, *Wiley, New-York.*

CONNIFFE D., (1985), Estimating regression equations with common explanatory variables but unequal numbers of observations, *Journal of Econometrics* 27, 179-196.

CONNIFFE D. and MORAN M.A., (1972), Double sampling with regression in comparative studies of carcass composition, *Biometrics* 28, 1011-1023.

COOK G.L., JONES D.W. and KEMPSTER A.J., (1983), A note on a simple criterion for choosing among sample joints for use in double sampling, *Animal Production* 36, 493-495.

ENGEL B. and WALSTRA P., (1991), Increasing precision or reducing expense in regression by using information from a concomitant variable, *Biometrics* 47, 13-20.

Annexe : non-biais et variance de $\hat{\beta}$

Les développements suivants visent à présenter les étapes principales du calcul de l'espérance et de la variance de $\hat{\beta}$. Une démonstration plus détaillée est disponible auprès de l'auteur.

Soit, pour tout $i \in \{0,1,2,\ldots,q\}$, les paramètres $\beta_{y|i} \in \mathbf{R}^p$, $\gamma_{y|i} \in \mathbf{R}^i$ et $\sigma^2_{y|i} \in \mathbf{R}^+$ tels que :

$$E\left[Y|Z^{(i)}\right] = X\beta_{y|i} + Z^{(i)}\gamma_{y|i}$$
$$V\left[Y|Z^{(i)}\right] = \sigma^2_{y|i},\ i = 0,1,\ldots,q.$$

En particulier, pour $i = 0$, $\beta_{y|0} = \beta$, $\gamma_{y|0} = 0$ et $\sigma^2_{y|0} = \sigma^2$ et pour $i = q$, $\beta_{y|q} = \beta_{y|z}$, $\gamma_{y|q} = \gamma_{y|z}$ et $\sigma^2_{y|q} = \sigma^2_{y|z}$.

On montre par récurrence décroissante sur i que l'estimateur du maximum de vraisemblance $\left(\hat{\beta}'_{y|i} \quad \hat{\gamma}'_{y|i}\right)'$ de $\left(\beta'_{y|i} \quad \gamma'_{y|i}\right)'$, $i = 0,1,\ldots,q$, est non-biaisé et de variance :

$$V[\begin{pmatrix}\hat{\beta}_{y|i}\\ \hat{\gamma}_{y|i}\end{pmatrix}|Z^{(i)}] = \sum_{j=i+1}^{q+1} \alpha^i_j \left[\left\{\left[X|Z^{(i)}\right]^{(n_j)}\right\}' \left[X|Z^{(i)}\right]^{(n_j)}\right]^{-1},\ i = 0,1,\ldots,q, \tag{13}$$

où $\alpha^i = \left(\alpha^i_{i+1} \quad \alpha^i_{i+2} \quad \ldots \quad \alpha^i_{q+1}\right)' \in \mathbf{R}^{q-i+1}$, $i = 0,1,\ldots,q$, est une suite de vecteurs définie par récurrence décroissante de la façon suivante :

$$\alpha^q = \sigma^2_{y|q},$$
$$\alpha^i = u_i + M_i\alpha^{i+1},$$

avec $u_i = \left(\sigma^2_{y|i} - \sigma^2_{y|i+1} \quad 0 \quad \ldots \quad 0\right)' \in \mathbf{R}^{q-i+1}$, $i = 0,1,\ldots,q-1$, et $M_i \in \mathcal{M}_{q-i+1,q-i}$ $(\mathbf{R})$ est définie de la façon suivante :

$$M_i = \begin{pmatrix} -\frac{1}{n_{i+2}-p-i-3} & -\frac{1}{n_{i+3}-p-i-3} & \ldots & -\frac{1}{n_{q+1}-p-i-3} \\ \frac{n_{i+2}-p-i-2}{n_{i+2}-p-i-3} & 0 & \ldots & 0 \\ 0 & \frac{n_{i+3}-p-i-2}{n_{i+3}-p-i-3} & \ldots & 0 \\ 0 & 0 & \ldots & \frac{n_{q+1}-p-i-2}{n_{q+1}-p-i-3} \end{pmatrix},$$

$i = 0,1,\ldots,q-1,$

et $\left[X|Z^{(i)}\right]^{(n_j)}$ est la matrice à n_j lignes et $p+i+1$ colonnes[1] obtenue en complétant la matrice $X^{(n_j)}$ des mesures des variables prédictrices sur le sous-échantillon de taille n_j par les colonnes de la matrice $\left(Z^{(i)}\right)^{(n_j)}$ des mesures des i premières variables auxiliaires sur le même sous-échantillon.

[1] la matrice X contient dans ce cas le vecteur dont toutes les composantes valent 1

Dans le cas où $i = q$, $\left(\hat{\beta}'_{y|q} \quad \hat{\gamma}'_{y|q}\right)'$ est obtenu par la méthode des moindres carrés ordinaires sur l'échantillon de taille n_{q+1}. Il est par conséquent non-biaisé conditionnellement à Z et

$$V\left[\begin{pmatrix}\hat{\beta}_{y|q}\\ \hat{\gamma}_{y|q}\end{pmatrix}|Z\right] = \sigma^2_{y|q}\left(\left\{[X|Z]^{(n_{q+1})}\right\}' [X|Z]^{(n_{q+1})}\right)^{-1},$$

où $[X|Z]^{(n_{q+1})}$ est la matrice à n_{q+1} lignes et $p+q+1$ colonnes[1] obtenue en complétant la matrice $X^{(n_{q+1})}$ des mesures des variables prédictrices sur le sous-échantillon de taille n_{q+1} par les colonnes de la matrice $Z^{(n_{q+1})}$ des mesures des variables auxiliaires sur le même sous-échantillon.

Le postulat (13) est donc vrai au rang q. Le passage du rang $i+1$ au rang i correspond à l'introduction d'une nouvelle variable auxiliaire et d'un nouveau sous-échantillon. Le calcul de l'espérance et de la variance s'appuie donc sur les techniques utilisées par Engel et Walstra (1991) dans le contexte univariable. On en déduit en particulier que $\left(\hat{\beta}'_{y|i} \quad \hat{\gamma}'_{y|i}\right)'$ est non-biaisé conditionnellement à $Z^{(i)}$ et que sa variance s'obtient de la façon suivante :

$$\begin{aligned}
&V\left[\begin{pmatrix}\hat{\beta}_{y|i}\\ \hat{\gamma}_{y|i}\end{pmatrix}|Z^{(i)}\right]\\
&= (\sigma^2_{y|i} - \sigma^2_{y|i+1} - \sum_{j=i+2}^{q+1} \frac{\alpha_j^{i+1}}{n_j - p - i - 3})\left(\left\{\left[X|Z^{(i)}\right]^{(n_{i+1})}\right\}' \left[X|Z^{(i)}\right]^{(n_{i+1})}\right)^{-1}\\
&\quad + \sum_{j=i+2}^{q} \alpha_j^{i+1} \frac{n_j - p - i - 2}{n_j - p - i - 3}\left(\left\{\left[X|Z^{(i)}\right]^{(n_j)}\right\}' \left[X|Z^{(i)}\right]^{(n_j)}\right)^{-1}.
\end{aligned}$$

On montre que le postulat de récurrence est vrai au rang i par identification de la relation précédente et de l'expression (13).