

Une note sur la statistique C_p sous un modèle de régression à erreur emboîtée

Jane L. Meza et P. Lahiri ¹

Résumé

Les modèles de régression à erreur emboîtée sont utilisés fréquemment pour l'estimation par petits domaines et les problèmes connexes. Cependant, l'application des critères standard de sélection du modèle de régression aux modèles à erreur emboîtée donne parfois lieu à des méthodes de sélection du modèle inefficaces. Nous illustrons ce point en examinant les propriétés de la statistique C_p au moyen d'une étude par simulation de Monte Carlo. L'inefficacité de la statistique C_p peut, cependant, être corrigée grâce à une transformation appropriée des données.

Mots clés : Statistiques C_p ; modèle de régression à erreur emboîtée; simulation de Monte Carlo.

1. Introduction

Nous examinons les limites d'un critère de sélection standard du modèle de régression, c'est-à-dire la statistique C_p , quand on l'applique au modèle de régression à erreur emboîtée. La statistique C_p (Mallows 1973) est définie par

$$C_p = \frac{SCR_p}{\hat{\sigma}^2} - n + 2p \quad (1)$$

où SCR_p est la somme des carrés des résidus et p est le nombre de paramètres du modèle P , n est le nombre d'observations et $\hat{\sigma}^2$ est une estimation de σ^2 . Si le modèle est correct, la valeur de C_p doit être semblable ou inférieure à p . Le critère de sélection du modèle C_p est sensible aux valeurs aberrantes et aux écarts par rapport à l'hypothèse d'erreurs i.i.d. suivant une loi normale. La statistique C_p ne peut, par conséquent, être appliquée directement au modèle de régression à erreur emboîtée, pour lequel la structure de l'erreur n'est pas i.i.d.

Nous proposons une transformation des données qui corrige la corrélation intragrupes et permet d'utiliser le critère standard de sélection du modèle C_p . La méthode que nous présentons ici peut être appliquée pour choisir des covariables dans l'analyse des données d'enquête complexes et aux modèles d'estimation par petits domaines. Par exemple, elle pourrait être utilisée pour sélectionner les covariables dans le modèle de régression à erreur emboîtée utilisé par Battese, Harter et Fuller (1988) pour estimer la superficie (en hectares) des cultures de maïs ou de soja pour douze comtés de l'Iowa. Ces auteurs ont utilisé le modèle suivant :

$$y_{ij} = x'_{ij} \beta + v_i + e_{ij}, \quad (2)$$

pour l'unité $j = 1, \dots, n_i$ dans le comté $i = 1, \dots, m$, où n_i est la taille de l'échantillon pour le petit domaine i et la taille totale de l'échantillon est $n = \sum_{i=1}^m n_i$. Les effets de comté, v_i , suivent une loi $N(0, \sigma_v^2)$ indépendante des erreurs aléatoires e_{ij} , qui suivent une loi $N(0, \sigma_e^2)$. La superficie (en hectares) dans l'unité j du comté i est dénotée y_{ij} et $x_{ij} = (1, x_{ij1}, \dots, x_{ijp})$ est un vecteur de dimension $p+1$ des valeurs des covariables x_1, \dots, x_p pour l'unité j dans le comté i . Le vecteur $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ est un vecteur de dimension $p+1$ de paramètres inconnus.

Le modèle de régression à erreur emboîtée peut être exprimé sous la forme matricielle suivante

$$y = X \beta + \varepsilon \quad (3)$$

où $y = (y'_1, \dots, y'_m)'$, $y'_i = (y_{i1}, \dots, y_{in_i})$, $\varepsilon = (\varepsilon'_1, \dots, \varepsilon'_m)'$, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})'$, $\varepsilon_{ij} = v_i + e_{ij}$. En outre, $X' = (X'_1, \dots, X'_m)$ où X_i est une matrice de dimensions $n_i \times (p+1)$ avec les lignes x_{ij} pour $j = 1, \dots, n_i$, $\varepsilon \sim N(0, \sigma^2 V)$ où $\sigma^2 = \sigma_v^2 + \sigma_e^2$, V a la forme d'une matrice diagonale par blocs $\bigoplus_1^m V_i$ avec $V_i = (1-\rho)I_{n_i} + \rho J_{n_i}$ où $\rho = \sigma_v^2 / \sigma^2$ est le coefficient de corrélation intrastrate courant, I_{n_i} est la matrice identité de dimensions $n_i \times n_i$ et J_{n_i} est la matrice unitaire de dimensions $n_i \times n_i$.

Puisque les erreurs du modèle à erreur emboîtée ne sont pas i.i.d., nous ne pouvons appliquer les procédures de régression standards. L'étude par simulation décrite à la section 3 montre que le critère C_p donne de mauvais résultats sous le modèle de régression à erreur emboîtée. Les transformations envisagées à la section suivante sont utilisées pour transformer le modèle de régression à erreur emboîtée en un modèle de régression standard à erreurs i.i.d. Appliqué à ces observations transformées, le critère C_p donne de nettement meilleurs résultats.

1. Jane L. Meza, University of Nebraska Medical Center, 984350 Nebraska Medical Center, Omaha, NE 68198-4350. Courriel : jmeza@unmc.edu; P. Lahiri, University of Maryland at College Park, 1218 Le Frak Hall, College Park, MD 20742-8241. Courriel : Plahiri@survey.umd.edu.

2. Correction pour les corrélations intradomaines

Comme nous l'avons mentionné à la section précédente, les méthodes classiques de sélection du modèle, telles que l'application du critère C_p , ne conviennent pas, puisqu'elles ne tiennent pas compte des corrélations intrastrates. Wu, Holt et Holmes (1988), ainsi que Rao, Sutradhar et Yue (1993) ont étudié l'effet des méthodes classiques dans le cas du modèle de régression à erreur emboîtée dans un contexte différent.

Considérons le modèle de régression à erreur emboîtée et posons que $\sigma^2 = \sigma_v^2 + \sigma_e^2$ et que ρ est le coefficient de corrélation intradomaine ordinaire $\rho = \sigma_v^2 / \sigma^2$. Comme dans Fuller et Battese (1973) et dans Rao et coll. (1993), transformons le modèle de régression à erreur emboîtée en un modèle de régression standard avec erreur i.i.d.

Soit

$$\alpha_i = 1 - \left[\frac{1 - \rho}{1 + (n_i - 1)\rho} \right]^{1/2}, \quad (4)$$

$$y_{ij}^* = y_{ij} - \alpha_i \bar{y}_i, \quad (5)$$

$$x_{ij}^* = x_{ij} - \alpha_i \bar{x}_i, \quad (6)$$

où $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ et $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij} / n_i$. Le modèle transformé devient alors

$$y_{ij}^* = x_{ij}^* \beta + e_{ij}^*, \quad (7)$$

pour $j = 1, \dots, n_i, i = 1, \dots, m$ et les e_{ij}^* sont indépendantes et de même loi $N(0, \sigma_e^2)$. Maintenant, nous pouvons appliquer le critère de sélection du modèle standard C_p aux données transformées.

En pratique, ρ est généralement inconnu et doit être estimé d'après les données. Rao et coll. (1993) ont utilisé la méthode de Henderson (1953) pour obtenir les estimateurs quadratiques sans biais $\hat{\sigma}_v^2$ et $\hat{\sigma}_e^2$ des composantes de la variance σ_v^2 et σ_e^2 . Une fois ces estimateurs obtenus, $\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$ peut être estimé par

$$\hat{\rho} = \max \left[0, \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2} \right]. \quad (8)$$

Pour obtenir les estimateurs des composantes de la variance, représentons par $\{u_{ij}\}$ les résidus de la régression par les moindres carrés ordinaires de $\{y_{ij} - \bar{y}_i\}$ sur $\{x_{ij1} - \bar{x}_{i,1}, \dots, x_{ijp} - \bar{x}_{i,p}\}$ sans le terme d'ordonnée à l'origine, où $\bar{x}_{i,l} = \sum_{j=1}^{n_i} x_{ijl} / n_i$ pour $l = 1, \dots, p$. Soit $\{r_{ij}\}$ les résidus de la régression par les moindres carrés ordinaires de y_{ij} sur $\{x_{ij0}, \dots, x_{ijp}\}$ avec le terme d'ordonnée à l'origine.

Les estimateurs de σ_v^2 et σ_e^2 sont donnés par

$$\hat{\sigma}_e^2 = (n - m - p - 1 - \lambda)^{-1} \sum_{i=1}^m \sum_{j=1}^{n_i} e_{ij}^2, \quad (9)$$

$$\hat{\sigma}_v^2 = n_*^{-1} \left[\sum_{i=1}^m \sum_{j=1}^{n_i} r_{ij}^2 - (n - p - 1) \hat{\sigma}_e^2 \right], \quad (10)$$

$$n_* = n - \text{tr} \left[(X'X)^{-1} \sum_{i=1}^m n_i^2 \bar{x}_i \bar{x}_i' \right] \quad (11)$$

où $\lambda = 0$ si le modèle ne contient pas de terme d'ordonnée à l'origine et $\lambda = 1$ autrement. Nous proposons d'appliquer le critère standard de sélection du modèle C_p à ces observations transformées y_{ij}^* et x_{ij}^* .

3. Une étude par simulation

Nous avons réalisé une étude par simulation pour examiner le comportement du critère de sélection du modèle C_p et des transformations proposées pour le modèle de régression à erreur emboîtée. Nous avons considéré le modèle suivant :

$$y_{ij} = \beta_0 x_{ij0} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + v_i + e_{ij} \quad (12)$$

pour $i = 1, \dots, 10, n_i \in \{2, \dots, 5\}, j = 1, \dots, n_i$ et $n = 40$. Les v_i suivent une loi $N(0, \sigma_v^2)$ indépendante des e_{ij} qui suivent une loi $N(0, 1)$. Les données x_{ijl} sont tirées d'un exemple donné par Gunst et Mason (1980) et inclus dans Shao (1993) (tableau 1). La valeur de x_{ij0} est 1 pour tous $i = 1, \dots, 10, j = 1, \dots, n_i$.

Comme certains coefficients β_k peuvent être nuls, nous avons choisi, à partir de $(x_0, x_1, x_2, x_3, x_4)$, diverses combinaisons de variables comme prédicteurs pour générer les données provenant d'un modèle de régression à erreur emboîtée. Il existe $2^p - 1 = 31$ modèles possibles. Chacun est dénoté par un sous-ensemble de $(0, 1, 2, 3, 4)$ qui contient les indices des variables x_i qui y sont incluses.

Pour générer les données, nous avons exécuté 1 000 simulations pour plusieurs valeurs de σ_v^2 afin d'estimer la probabilité de sélection de chaque modèle au moyen du critère C_p . Nous avons donné la valeur 1 à σ_e^2 pour toutes les simulations. Les résultats des simulations sont présentés au tableau 2. Nous avons considéré les valeurs 0, 1, 2, 5, 10 et 16 pour σ_v^2 et fixé les valeurs de β' à $(2, 0, 0, 4, 0), (2, 0, 0, 4, 8), (2, 9, 0, 4, 8)$ et $(2, 9, 6, 4, 8)$ comme dans Shao (1993). Les modèles ont été répartis en trois catégories, à savoir optimal, catégorie II (correct mais non optimal) ou catégorie I (incorrect).

Le critère C_p a donné de mauvais résultats pour les grandes valeurs de σ_v^2 . Pour le modèle $\beta' = (2, 0, 0, 4, 0)$ avec $\sigma_v^2 = 1$, les probabilités de sélection estimées étaient : modèle optimal, 0,54; modèle correct, 0,46; modèle incorrect, 0. Par contre, pour $\sigma_v^2 = 16$, les probabilités de

Tableau 1
Données pour la simulation de l'erreur emboîtée

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0,3600	0,5300	1,0600	0,5326	0,0900	0,1800	0,5900	0,1855
1,3200	2,5200	5,7400	3,6183	0,0200	0,1600	0,2400	0,1572
0,0600	0,0900	0,2700	0,2594	0,0200	0,1100	0,2100	0,0998
0,1600	0,4100	0,8300	1,0346	0,0500	0,2400	0,4300	0,2804
0,0100	0,0200	0,0700	0,0381	0,1100	0,3900	0,2900	0,2879
0,0200	0,0700	0,0700	0,3440	0,1800	0,1100	0,4300	0,6810
0,5600	0,6200	2,1200	1,4559	0,0400	0,0900	0,2300	0,3242
0,9800	1,0600	2,8900	4,0182	0,8500	1,3300	2,7000	2,6013
0,3200	0,2000	0,7600	0,4600	0,1700	0,3200	0,6600	0,4469
0,0100	0,0000	0,0700	0,1540	0,0800	0,1200	0,4900	0,2436
0,1500	0,2500	0,5000	0,6516	0,3800	0,1800	0,4900	0,4400
0,2400	0,2800	0,5900	0,0611	0,1100	0,1300	0,1800	0,3351
0,1100	0,3500	0,4000	0,1922	0,3900	0,3800	0,9900	1,3979
0,0800	0,1300	0,2800	0,0931	0,4300	0,4600	1,4700	2,0138
0,6100	0,8500	0,4900	0,0538	0,5700	1,1600	1,8200	1,9356
0,0300	0,0300	0,2300	0,0199	0,1300	0,0300	0,0800	0,1050
0,0600	0,1100	0,5000	0,0419	0,0400	0,0500	0,1400	0,2207
0,0200	0,0800	0,2500	0,1093	0,1300	0,1800	0,2800	0,0180
0,0400	0,2400	0,0800	0,0328	0,2000	0,9500	0,4100	0,1017
0,0000	0,0200	0,0400	0,0797	0,0700	0,0600	0,1800	0,0962

sélection estimées étaient : modèle optimal, 0,43; modèle correct, 0,35; modèle incorrect, 0,22. Le critère C_p n'a pas donné de bons résultats non plus pour les modèles plus grands avec de grandes valeurs de σ_v^2 . En revanche, il a donné de très bons résultats pour les grands modèles avec de petites valeurs de σ_v^2 . Pour le modèle complet $\beta' = (2, 9, 6, 4, 8)$, avec $\sigma_v^2 = 1$, les probabilités de sélection estimées étaient : modèle optimal, 0,98; modèle correct, 0,02; modèle incorrect, 0. Comparativement, pour $\sigma_v^2 = 16$, les probabilités de sélection estimées étaient : modèle optimal, 0,11; modèle incorrect, 0,89. Il convient de souligner que, dans ce scénario, le seul modèle correct est le modèle optimal.

En résumé, quand on applique le critère C_p à des données obéissant au modèle de régression à erreur emboîtée :

1. pour n'importe quel modèle, la probabilité estimée de sélection du modèle *optimal* diminue quand σ_v^2 augmente;
2. pour n'importe quel modèle, la probabilité estimée de sélection d'un modèle *incorrect* augmente quand σ_v^2 augmente;
3. à mesure que le nombre de variables incluses dans le modèle augmente et que σ_v^2 augmente, la probabilité estimée de sélection du modèle *optimal* diminue;
4. à mesure que le nombre de variables incluses dans le modèle augmente et que σ_v^2 augmente, la probabilité estimée de sélection d'un modèle *incorrect* augmente.

Nous avons alors utilisé les données pour estimer la probabilité de sélectionner chaque modèle en utilisant le critère C_p sous la transformation pour un coefficient de corrélation ρ connu. Les résultats de la simulation sont donnés dans le tableau 3. Pour le modèle $\beta' = (2, 0, 0, 4, 0)$ avec $\sigma_v^2 = 0$ (modèle de régression standard), les probabilités de sélection estimées étaient : modèle optimal, 0,62; modèle correct, 0,38; modèle incorrect, 0 (tableau 2). Pareillement, sous la transformation pour ρ connu avec $\sigma_v^2 = 16$, les probabilités de sélection estimées étaient : modèle optimal, 0,60; modèle correct, 0,40; modèle incorrect, 0 (tableau 3). Pour le modèle complet $\beta' = (2, 9, 6, 4, 8)$, la probabilité estimée de sélectionner le modèle optimal était 1 pour le modèle de régression standard (tableau 2, $\sigma_v^2 = 0$), ainsi que sous la transformation avec ρ connu pour toutes les valeurs de σ_v^2 envisagées (tableau 3).

En pratique, le coefficient de corrélation ρ est inconnu et doit être estimé d'après les données. Par conséquent, la transformation est plus utile pour les praticiens quand ρ est inconnu. Les résultats de la transformation dans ces conditions sont présentés au tableau 4. Quand nous avons estimé ρ , la probabilité estimée de sélectionner le modèle optimal ou un modèle correct n'a diminué que légèrement. La diminution la plus importante de la probabilité estimée de sélectionner le modèle optimal était 0,03 pour le modèle avec $\beta' = (2, 0, 4, 0)$ et $\sigma_v^2 = 1$, soit 0,61 pour ρ connu (tableau 3) comparativement à 0,58 pour ρ inconnu (tableau 4).

Tableau 2
Probabilités de sélection du modèle avant transformation

$\beta = (2, 0, 0, 4, 0)'$							
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0,62	0,54	0,49	0,46	0,45	0,43
0, 2, 3	II	0,11	0,09	0,09	0,10	0,07	0,06
0, 1, 3	II	0,09	0,14	0,19	0,17	0,15	0,12
0, 3, 4	II	0,09	0,13	0,13	0,14	0,11	0,10
0, 1, 2, 3	II	0,03	0,05	0,06	0,05	0,04	0,04
0, 1, 3, 4	II	0,02	0,03	0,02	0,02	0,02	0,01
0, 2, 3, 4	II	0,02	0,01	0,02	0,02	0,01	0,02
0, 1, 2, 3, 4	II	0,02	0,01	0,00	0,00	0,01	0,00
0, 1	I	0,00	0,00	0,00	0,01	0,07	0,09
0, 2	I	0,00	0,00	0,00	0,01	0,03	0,05
0, 4	I	0,00	0,00	0,00	0,00	0,01	0,04
0, 1, 2	I	0,00	0,00	0,00	0,01	0,01	0,01
0, 1, 4	I	0,00	0,00	0,00	0,01	0,02	0,03
0, 1, 2, 4	I	0,00	0,00	0,00	0,00	0,00	0,00
$\beta = (2, 0, 0, 4, 8)'$							
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0,72	0,67	0,63	0,61	0,58	0,49
0, 2, 3, 4	II	0,12	0,12	0,14	0,14	0,11	0,09
0, 1, 3, 4	II	0,12	0,16	0,18	0,14	0,12	0,11
0, 1, 2, 3, 4	II	0,04	0,05	0,05	0,05	0,04	0,04
0, 4	I	0,00	0,00	0,00	0,00	0,01	0,06
0, 1, 4	I	0,00	0,00	0,00	0,02	0,05	0,10
0, 2, 4	I	0,00	0,00	0,00	0,03	0,07	0,10
0, 1, 2, 4	I	0,00	0,00	0,00	0,00	0,01	0,01
$\beta = (2, 9, 0, 4, 8)'$							
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0,83	0,78	0,75	0,63	0,39	0,25
0, 1, 2, 3, 4	II	0,17	0,20	0,18	0,13	0,09	0,07
0, 3, 4	I	0,00	0,01	0,03	0,13	0,29	0,35
0, 1, 4	I	0,00	0,00	0,00	0,03	0,11	0,15
0, 2, 3, 4	I	0,00	0,01	0,03	0,07	0,06	0,09
0, 2, 4	I	0,00	0,00	0,00	0,00	0,02	0,05
0, 1, 2, 4	I	0,00	0,00	0,00	0,02	0,04	0,04
$\beta = (2, 9, 6, 4, 8)'$							
Modèle	Catégorie	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1,00	0,98	0,90	0,60	0,29	0,11
0, 2, 3, 4	I	0,00	0,02	0,07	0,24	0,32	0,28
0, 1, 3, 4	I	0,00	0,00	0,02	0,11	0,18	0,23
0, 1, 2, 4	I	0,00	0,00	0,01	0,06	0,13	0,17
0, 3, 4	I	0,00	0,00	0,00	0,00	0,03	0,09
0, 2, 4	I	0,00	0,00	0,00	0,00	0,03	0,10
0, 1, 4	I	0,00	0,00	0,00	0,00	0,01	0,03
0, 1, 3	I	0,00	0,00	0,00	0,00	0,00	0,00

Tableau 3
Probabilités de sélection du modèle après transformation, ρ connu

$\beta = (2, 0, 0, 4, 0)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0,61	0,60	0,61	0,61	0,60
0, 3, 4	II	0,11	0,10	0,11	0,11	0,11
0, 2, 3	II	0,10	0,11	0,11	0,10	0,11
0, 1, 3	II	0,09	0,10	0,08	0,09	0,09
0, 1, 2, 3	II	0,04	0,04	0,04	0,04	0,04
0, 1, 3, 4	II	0,03	0,03	0,03	0,02	0,02
0, 2, 3, 4	II	0,02	0,02	0,02	0,02	0,02
0, 1, 2, 3, 4	II	0,01	0,01	0,01	0,01	0,01
$\beta = (2, 0, 0, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0,71	0,71	0,73	0,72	0,71
0, 2, 3, 4	II	0,13	0,12	0,11	0,12	0,13
0, 1, 3, 4	II	0,11	0,12	0,10	0,11	0,11
0, 1, 2, 3, 4	II	0,05	0,05	0,05	0,05	0,05
$\beta = (2, 9, 0, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0,82	0,83	0,83	0,82	0,83
0, 1, 2, 3, 4	II	0,18	0,17	0,17	0,18	0,17
$\beta = (2, 9, 6, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1,00	1,00	1,00	1,00	1,00

Tableau 4
Probabilités de sélection du modèle après transformation, ρ inconnu

$\beta = (2, 0, 0, 4, 0)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3	Optimal	0,58	0,59	0,60	0,61	0,60
0, 3, 4	II	0,11	0,10	0,11	0,10	0,10
0, 2, 3	II	0,11	0,10	0,11	0,11	0,11
0, 1, 3	II	0,08	0,09	0,10	0,09	0,09
0, 1, 2, 3	II	0,04	0,04	0,03	0,04	0,04
0, 1, 3, 4	II	0,03	0,03	0,02	0,02	0,02
0, 2, 3, 4	II	0,03	0,03	0,02	0,02	0,03
0, 1, 2, 3, 4	II	0,02	0,02	0,01	0,01	0,01
$\beta = (2, 0, 0, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 3, 4	Optimal	0,70	0,70	0,70	0,71	0,70
0, 2, 3, 4	II	0,13	0,14	0,13	0,13	0,13
0, 1, 3, 4	II	0,13	0,11	0,12	0,11	0,12
0, 1, 2, 3, 4	II	0,04	0,05	0,05	0,05	0,05
$\beta = (2, 9, 0, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 3, 4	Optimal	0,82	0,82	0,81	0,83	0,83
0, 1, 2, 3, 4	II	0,18	0,18	0,19	0,17	0,17
$\beta = (2, 9, 6, 4, 8)'$						
Modèle	Catégorie	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1,00	1,00	1,00	1,00	1,00

D'après les résultats de nos simulations, quand le critère de sélection C_p est appliqué à des données obéissant au modèle de régression erreur emboîtée :

1. sous les deux transformations (ρ connu et ρ inconnu), la probabilité estimée de sélectionner un modèle *incorrect* est 0;
2. sous la transformation pour ρ connu, la probabilité de sélectionner le modèle *optimal* est semblable à celle du modèle de régression standard;
3. quand on doit estimer ρ , la probabilité estimée de sélectionner le modèle optimal ou un modèle correct ne diminue que légèrement;
4. sous les deux transformations (ρ connu et ρ estimé), le critère C_p donne de bons résultats, même pour des modèles plus grands avec grande valeur de σ_v^2 ;
5. les propriétés du critère C_p pour le modèle de régression à erreur emboîtée ressemblent à celles du critère C_p pour le modèle de régression standard.

En résumé, le critère C_p donne de mauvais résultats sous le modèle de régression à erreur emboîtée quand la valeur de σ_v^2 est grande. Quand on applique la transformation pour ρ inconnu (ou ρ connu), le modèle devient un modèle de régression standard et la statistique C_p se comporte en conséquence.

Remerciements

L'étude a été financée partiellement par une bourse de l'organisation Gallup.

Bibliographie

- Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- Fuller, W.A., et Battese, G.E. (1973). Transformations for estimation of linear models with nested error structures. *Journal of the American Statistical Association*, 68, 626-632.
- Gunst, G.F., et Mason, R.L. (1980). *Regression Analysis and Its Application*. New York : Marcel Dekker.
- Henderson, C.R. (1953). Estimation of variance and variance components. *Biometrics*, 9, 226-252.
- Mallows, C.L. (1973). Some comments on C_p . *Technometrics*, 15, 661-675.
- Rao, J.N.K., Sutradhar, B.C. et Yue, K. (1993). Generalized least squares F test in regression analysis with two-stage cluster samples. *Journal of the American Statistical Association*, 88, 1388-1391.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.
- Wu, C.F.J., Holt, D. et Holmes, D.J. (1988). The effect of two-stage sampling on the F Statistic. *Journal of the American Statistical Association*, 83, 150-159.