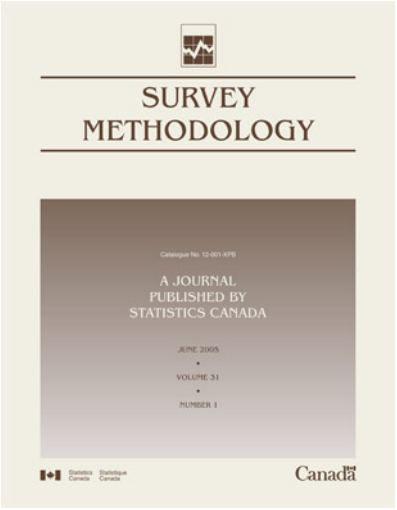
Catalogue no. 12-001-XIE

Survey Methodology

June 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 1, pp. 105-109
Statistics Canada, Catalogue No. 12-001

A Note on the C_p Statistic Under the Nested Error Regression Model

Jane L. Meza and P. Lahiri [1]

Abstract

Nested error regression models are frequently used in small-area estimation and related problems. Standard regression model selection criterion, when applied to nested error regression models, may result in inefficient model selection methods. We illustrate this point by examining the performance of the C_P statistic through a Monte Carlo simulation study. The inefficiency of the C_P statistic may, however, be rectified by a suitable transformation of the data.

Key Words: C_P statistics; Nester error regression model; Monte Carlo simulation.

1. Introduction

This paper examines the limitations of a standard regression model selection criterion, C_P the statistic, for nested error regression models. The C_P statistic (Mallows 1973) is defined by

$$C_P = \frac{\text{RSS}_P}{\hat{\sigma}^2} - n + 2p \tag{1}$$

where RSS_P is the residual sum of squares and p is the number of parameters for model P, n is the number of observations and $\hat{\sigma}^2$ is an estimate of σ^2. If the model is correct, the value of C_P should be similar to or smaller than p. The C_P model selection criterion is sensitive to outliers and departures from the normal i.i.d. assumption on the errors. The C_P statistic therefore cannot be directly applied to the nested error regression model since here the error structure is not i.i.d.

We propose a transformation that adjusts for intracluster correlation and allows use of the standard C_P model selection criterion. The method presented in this paper can be applied to select covariates in the analysis of complex survey data and small-area models. For example, our technique could be used to select covariates in the nested error regression model used by Battese, Harter and Fuller (1988) to estimate the area planted (in hectacres) with corn or soybeans for twelve Iowa counties. They used the following model:

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \tag{2}$$

for unit $j = 1, \ldots, n_i$ in county $i = 1, \ldots, m$, where n_i is the sample size for small area i and the total sample size is $n = \sum_{i=1}^{m} n_i$. The county effects, v_i, are distributed as $N(0, \sigma_v^2)$ independent of the random errors e_{ij}, which are distributed as $N(0, \sigma_e^2)$. The area (in hectacres) in unit j of county i is denoted by y_{ij} and $x_{ij} = (1, x_{ij1}, \ldots, x_{ijp})$ is a $p+1$ vector of the values of the covariates $x_1, \ldots, x_p$ for unit j in county i. The vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a $p+1$ vector of unknown parameters.

The nested error regression model can be expressed in matrix form as

$$y = X\beta + \varepsilon \tag{3}$$

where $y = (y_1', \ldots, y_m')'$, $y_i' = (y_{i1}, \ldots, y_{in_i})$, $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_m')'$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})'$, $\varepsilon_{ij} = v_i + e_{ij}$. Further, $\boldsymbol{X}' = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')$ where X_i is an $n_i \times (p+1)$ matrix with rows x_{ij} for $j = 1, \ldots, n_i$, $\varepsilon \sim N(0, \sigma^2 V)$ where $\sigma^2 = \sigma_v^2 + \sigma_e^2$, V has block-diagonal form $\oplus_1^m V_i$ with $V_i = (1-\rho) I_{n_i} + \rho J_{n_i}$ where $\rho = \sigma_v^2 / \sigma^2$ is the common instrastratum correlation, I_{n_i} is the $n_i \times n_i$ identity matrix and J_{n_i} is the $n_i \times n_i$ unit matrix.

Since the nested error model does not have i.i.d errors, standard regression procedures do not apply. The simulation study in section 3 reveals that the C_P criterion does not perform well under the nested error regression model. The transformations considered in the next section are used to transform the nested error regression model into a standard regression model with i.i.d. errors. With these transformed observations, the C_P criterion performs much better.

2. Adjusting for Intra-area Correlations

As noted in the previous section, conventional model selection methods like the C_P criterion are not appropriate since the instrastratum correlations are ignored. Wu, Holt and Holmes (1988) and Rao, Sutradhar and Yue (1993) studied the effect of conventional methods for the nested error regression model in a different context.

Consider the nested error regression model and let $\sigma^2 = \sigma_v^2 + \sigma_e^2$ and ρ be the common intra-area correlation, $\rho = \sigma_v^2 / \sigma^2$. As in Fuller and Battese (1973) and Rao *et al.*

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

(1993), transform the nested error regression model into a standard regression model with i.i.d. errors.

Let

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

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

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

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ and $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij} / n_i$. The transformed model then becomes

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

for $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ and e_{ij}^* are independently distributed as $N(0, \sigma_e^2)$. Now, the standard C_P model selection criterion may be applied to the transformed data.

In practice, ρ is usually unknown and must be estimated from the data. Rao *et al.* (1993) used Henderson's (1953) method to obtain unbiased quadratic estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ of the variance components σ_v^2 and σ_e^2. Once the estimators have been obtained, $\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$ may be estimated by

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

To obtain the estimators of the variance components, let $\{u_{ij}\}$ be the residuals from the ordinary least squares regression of $\{y_{ij} - \bar{y}_{i.}\}$ on $\{x_{ij1} - \bar{x}_{i.1}, \ldots, x_{ijp} - \bar{x}_{i.p}\}$ without the intercept term, where $x_{i.l} = \sum_{j=1}^{n_i} x_{ijl} / n_i$ for $l = 1, \ldots, p$. Let $\{r_{ij}\}$ be the residuals from the ordinary least squares regression of y_{ij} on $\{x_{ij0}, \ldots, x_{ijp}\}$ with the intercept term.

The estimators of σ_v^2 and σ_e^2 are given by

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

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

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

where $\lambda = 0$ if the model has no intercept term and $\lambda = 1$ otherwise. We propose to apply standard C_P model selection criterion on these transformed observations y_{ij}^* and x_{ij}^*.

3. A Simulation Study

A simulation study was conducted to examine the behavior of the C_P model selection criterion and the proposed transformations for the nested error regression model. The following model was considered:

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

for $i = 1, \ldots, 10$, $n_i \in \{2, \ldots, 5\}$, $j = 1, \ldots, n_i$ and $n = 40$. The v_i are distributed as $N(0, \sigma_v^2)$ independent of e_{ij} which are distributed as $N(0, 1)$. The data x_{ijl} are taken from an example given by Gunst and Mason (1980) and included in Shao (1993) (Table 1). The value of x_{ij0} is 1 for all $i = 1, \ldots, 10$, $j = 1, \ldots, n_i$.

Table 1
Data for Nested Error Simulation

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

Some of the β_k may be zero and thus various combinations of variables were chosen from $(x_0, x_1, x_2, x_3, x_4)$ to be the predictors used to generate data coming from a nested error regression model. There are $2^p - 1 = 31$ possible models. Each model will be denoted by a subset of (0, 1, 2, 3, 4) that contains the indices of the variables x_l in the model.

Data were generated using 1,000 simulations for several values of σ_v^2 to estimate the probability of selecting each model using the C_P criterion. The value of σ_e^2 was taken to be 1 for all simulations. The results of the simulation are given in Table 2. The values of σ_v^2 considered were 0, 1, 2, 5, 10 and 16 and the values of β' were taken to be (2, 0, 0, 4, 0), (2, 0, 0, 4, 8), (2, 9, 0, 4, 8) and (2, 9, 6, 4, 8) as in Shao (1993). Models were categorized as optimal, category II (correct but not optimal), or category I (incorrect).

The C_P criterion did not perform well for large values of σ_v^2. For the model $\beta' = (2, 0, 0, 4, 0)$ with $\sigma_v^2 = 1$ the estimated selection probabilities were: optimal model, 0.54; correct model, 0.46; incorrect model, 0. In contrast, when $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.43; correct model, 0.35; incorrect model, 0.22.

The C_P criterion also did not perform well for larger models with large values of σ_v^2. The C_P criterion however

Table 2
Probabilities of Model Selection Before Transformation

		$\beta = (2, 0, 0, 4, 0)'$					
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
		$\beta = (2, 0, 0, 4, 8)'$					
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
		$\beta = (2, 9, 0, 4, 8)'$					
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
		$\beta = (2, 9, 6, 4, 8)'$					
Model	Category	$\sigma_v^2 = 0$	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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

did very well for large models with small values of σ_v^2. For the full model $\beta' = (2, 9, 6, 4, 8)$ with $\sigma_v^2 = 1$, the estimated selection probabilities were: optimal model, 0.98; correct model, 0.02; incorrect model, 0. In contrast, when $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.11; incorrect model, 0.89. Note that in this scenario there are no correct models other than the optimal model.

In summary, when the C_P criterion is applied to data following the nested error regression model:

1. For any particular model, the estimated probability of selecting the *optimal* model decreases as σ_v^2 increases.
2. For any particular model, the estimated probability of selecting an *incorrect* model increases as σ_v^2 increases.
3. As the number of variables included in the model increases and σ_v^2 increases, the estimated probability of selecting the *optimal* model decreases.
4. As the number of variables included in the model increases and σ_v^2 increases, the estimated probability of selecting an *incorrect* model increases.

The data were then used to estimate the probability of selecting each model using the C_P criterion under the transformation for ρ known. The results of the simulation are given in Table 3. For the model $\beta' = (2, 0, 0, 4, 0)$ with $\sigma_v^2 = 0$ (standard regression model) the estimated selection probabilities were: optimal model, 0.62; correct model, 0.38; incorrect model, 0 (Table 2). Similarly, under the transformation for ρ known with $\sigma_v^2 = 16$, the estimated selection probabilities were: optimal model, 0.60; correct model, 0.40; incorrect model, 0 (Table 3). For the full model $\beta' = (2, 9, 6, 4, 8)$, the estimated probability of selecting the optimal model was 1 for both the standard regression model (Table 2, $\sigma_v^2 = 0$) and under the transformation for ρ known for all values of σ_v^2 considered (Table 3).

In practice, ρ is unknown and must be estimated from the data. The transformation for ρ unknown is therefore more helpful for practitioners. The results for the transformation with ρ unknown are displayed in Table 4. When ρ was estimated, there was only a small decrease in the estimated probability of selecting the optimal model or a correct model. The largest decrease in the estimated probability of selecting the optimal model was 0.03 for the model with $\beta' = (2, 0, 4, 0)$ and $\sigma_v^2 = 1, 0.61$ for ρ known (Table 3) compared to 0.58 for ρ unknown (Table 4).

Table 3
Probabilities of Model Selection After Transformation, ρ Known

$\beta = (2, 0, 0, 4, 0)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
$\beta = (2, 0, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
$\beta = (2, 9, 0, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
$\beta = (2, 9, 6, 4, 8)'$						
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	1.00	1.00	1.00	1.00

Table 4
Probabilities of Model Selection After Transformation, ρ Unknown

				$\beta = (2, 0, 0, 4, 0)'$		
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
				$\beta = (2, 0, 0, 4, 8)'$		
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
				$\beta = (2, 9, 0, 4, 8)'$		
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
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
				$\beta = (2, 9, 6, 4, 8)'$		
Model	Category	$\sigma_v^2 = 1$	$\sigma_v^2 = 2$	$\sigma_v^2 = 5$	$\sigma_v^2 = 10$	$\sigma_v^2 = 16$
0, 1, 2, 3, 4	Optimal	1.00	1.00	1.00	1.00	1.00

Based on our simulation results, when the C_P criterion is applied to data following the nested error regression model:

1. Under both transformations (ρ known and ρ unknown), the estimated probability of selecting an *incorrect* model was 0.
2. Under the transformation for ρ known, the probability of selecting the *optimal* model was similar to that of the standard regression model.
3. When ρ was estimated, there was only a small decrease in the estimated probability of selecting the optimal model or a correct model.
4. Under both transformations (ρ known and ρ estimated), the C_P criterion performed well, even for larger models with large values of σ_v^2.
5. The performance of the C_P criterion for the nested error regression model resembles that of the C_P criterion for the standard regression model.

In summary, the C_P criterion does not perform well under the nested error regression model when σ_v^2 is large. When the transformation for ρ unknown (or ρ known) is applied, the model then becomes a standard regression model and the C_P statistic performs accordingly.

Acknowledgements

The research was supported in part by a grant from the Gallup Organization.

References

Battese, G.E., Harter, R.M. and 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., and 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., and 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. and 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. and Holmes, D.J. (1988). The effect of two-stage sampling on the F Statistic. *Journal of the American Statistical Association*, 83, 150-159.