Statistics Canada Statistique Canada

Ottawa

503/3

ESTIMATION OF REGRESSION PARAMETERS FOR FINITE POPULATIONS: A MONTE - CARLO STUDY

WORKING PAPER NO. BSMD 85-008E

MIKE HIDIROGLOU
STATISTICS CANADA
FEBRUARY 1983

TABLE OF CONTENTS

1. INTRODUCTION

Regression analysis is a very widely used tool for analysing multivariate data. The analysis of multivariate data has recently been greatly aided by the development of computer packages. Users of such packages quite often ignore the assumptions that should be supported by the data sets they analyse. One of the major assumptions underlying regression analysis in computer packages is that the observations are independent. This crucial assumption is violated if the data has been collected using cluster or multistage sampling designs. The subsequent analyses using standard computer packages do not take this important consideration into account with the net effect being that the standard estimators for the variance of the estimates regression coefficients are likely to be serious underestimates. Test statistics and confidence regions based on those variance estimators are also badly affected.

The problem of multiple regression estimation in finite population sampling has been studied by Konijn (1962), Frankel (1971), Fuller (1975), Hartley and Sielken (1975), Holt, Smith and Winter (1980), Särndal (1978) and more recently by Scott and Holt (1982). These authors have pointed out to the dangers of using traditional computer packages and have provided some theory to handle the non-independence problem caused by multistage or cluster sampling. There is not much literature on the applications of these theories to data sets. Some indication of the performance of the estimators of variance for the regression coefficients, using this theory has been reported by Frankel (1971), Shah et al. (1977), and Scott and Holt (1982). Frankel (1971) studied the empirical behavior of multiple regression coefficients computed from a stratified cluster sample. The data used for this study were a sample of U.S. households selected by the U.S. Bureau of the Census in the March 1967 Current Population Survey. The objective of the regression analysis was the estimation of the finite population parameters as defined by the population moments. Frankel used the Taylor approximations to the variance formula suggested by Tepping (1968).

In this paper, small sample properties of regression estimators and their estimated variance will be presented in the context of finite population sampling. That is, stratification and clustering will be taken into account when estimating the finite population regression parameters. Two types of regression procedures will be studied. One where the data are not subject to measurement error and the other where the data is subject to measurement error. The procedures for data with measurement error should be of particular interest to survey samplers because the data collected in sample surveys, particularly those collected from human respondents, are subject to measurement error. The U.S. Bureau of the Census (1972) has reported estimates of the response variance, as a percentage of the total variance, that range from 0.5 to 40 percent. Regression analyses performed under these circumstances must therefore take these errors into account. Fuller (1981) has extensively investigated the properties of measurement error (error in variables) for estimators of regression parameters.

The simulation was carried out using the computer program SUPER CARP (1981). The structure of the paper is as follows. The investigated models are presented in Section 2. The design of the sampling experiments and the simulation results are given in Sections 3 and 4 respectively.

2. MODELS

The finite population model is given by

$$\underset{\sim}{y}_N = \underset{\sim}{x}_N \underset{\sim}{B}_{OLS} + \underset{\sim}{e}_N \tag{2.1}$$

where $\underset{\sim}{y}_N$ is an N×1 vector of observations on the dependent variable; $\underset{\sim}{x}_N$ is an N×p non-stochastic matrix of observations on p regression; $\underset{\sim}{B}_{OLS}$ is the p-dimensional vector of regression coefficients; $\underset{\sim}{e}_N$ represents an N×1 vector of deviations from the linear relationship.

N is the size of the population interest. In the absence of measurement error on $\tilde{y}$ and $\tilde{x}$, minimizing the sum of the squared deviations over the entire population yields the following definition (Frankel, 1971) for the population regression coefficients $\tilde{B}_{OLS}$ as

$$\tilde{B}_{OLS} = (\tilde{x}_N^T \tilde{x}_N)^- \tilde{x}_N^T \tilde{y}_N \qquad (2.2)$$

where the inverse of $(\tilde{x}_N^T \tilde{x}_N)$ may be the Moore-Penrose inverse.

The estimator for $\tilde{B}_{OLS}$ is obtained via a two-stage stratified clustered sample obtained as follows. The population is first divided into $h = 1, 2, \ldots, L$ strata. For each stratum, a sample of size n_h is drawn from N_h clusters and from each selected cluster of size M_{hj} a sample of m_{hj} elements is drawn using a given drawing mechanism at each stage. A natural estimator for $\tilde{B}_{OLS}$ which appeals to survey statisticians is one which takes the survey weights into account. Such an estimator is given by

$$\tilde{b}_{OLS} = (\tilde{x}_n^T \tilde{W}_n \tilde{x}_n)^- \tilde{x}_n^T \tilde{W}_n \tilde{y}_n \qquad (2.3)$$

with the rs-th element of $\tilde{x}_n^T \tilde{W}_n \tilde{x}_n$ is given by

$$\sum_{h=1}^{L} \sum_{j=1}^{n_h} \sum_{k=1}^{m_{hj}} x_{hjkr}\, x_{hjks}\, w_{hjk}$$

where

w_{hjk} = weight associated with the hjk-th observation,

x_{hjkr} = the hjk-th observation on the r-th independent variables,

$$n = \sum_{h=1}^{L} \sum_{j=1}^{n_h} m_{hj} \text{ (the effective sample size).}$$

Similarly, the rs-th element of $\tilde{x}_n^T \tilde{W}_n \tilde{y}_n$ is given by

$$\sum_{h=1}^{L} \sum_{j=1}^{n_h} \sum_{k=1}^{m_{hj}} x_{hjks}\, y_{hjk}\, w_{hjk}$$

with y_{hjk} being the hjk-th observation on the dependent variable. Fuller (1975) provided an estimator for the covariance matrix of $\underset{\sim}{b}_{OLS}$ under certain regularity conditions to ensure the convergence to normality and consistency. This estimator is given by

$$\hat{\underset{\sim}{V}}_{OLS} = (\underset{\sim}{x}_n^T \underset{\sim}{W}_n \underset{\sim}{x}_n)^- \hat{\underset{\sim}{G}}_{OLS} (\underset{\sim}{x}_n^T \underset{\sim}{W}_n \underset{\sim}{x}_n)^- \tag{2.4}$$

where the rs-th element of $\hat{\underset{\sim}{G}}_{OLS}$ is

$$g_{rs} = \sum_{h=1}^{L} \frac{n_h(1-f_h)}{n_h-1} \sum_{j=1}^{n_h} (\hat{d}_{hj.r} - \bar{d}_{h..r})(\hat{d}_{hj.s} - \bar{d}_{h..s}) \tag{2.5}$$

with $$f_h = n_h/N_h$$

$$\hat{d}_{hjkr} = x_{hjkr}\, \hat{v}_{hjk}\, w_{hjk} ,$$

$$\hat{v}_{hjk} = y_{hjk} - \sum_{r=1}^{p} b_{OLS,r}\, x_{hjkr} ,$$

$$d_{hj.r} = \sum_{k=1}^{m_{hj}} \hat{d}_{hjkr} ,$$

and

$$\bar{d}_{h..r} = \sum_{j=1}^{n_h} \hat{d}_{hj.r}/n_h \quad .$$

In the presence of measurement error, the finite population model is given by

$$\underset{\sim}{Y}_N = \underset{\sim}{X}_N \underset{\sim}{B}_{EV} + \underset{\sim}{e}_N \tag{2.6}$$

where $\underset{\sim}{Y}_N$ and $\underset{\sim}{X}_N$ are the observed random variables incorporating

measurement error. That is

$$\underset{\sim}{X}_N = \underset{\sim}{x}_N + \underset{\sim}{u}_N \text{ and } \underset{\sim}{Y}_N = \underset{\sim}{y}_N + \underset{\sim}{\varepsilon}_N$$

where $\underset{\sim}{\delta}_N = (\underset{\sim}{\varepsilon}_N, \underset{\sim}{u}_N)$ is the matrix of response errors for the population. Assume that the covariance matrix for $\underset{\sim}{u}_N$ is known and is denoted as $\underset{\sim}{\Sigma}_{uu}$. Similarly to the ordinary least square case ($\underset{\sim}{B}_{OLS}$), $\underset{\sim}{B}_{EV}$ may be defined as

$$\underset{\sim}{B}_{EV} = (\underset{\sim}{X}_N^T \underset{\sim}{X}_N - N \underset{\sim}{\Sigma}_{uu})^{-} \underset{\sim}{X}_N^T \underset{\sim}{Y}_N . \tag{2.7}$$

A sample estimator for $\underset{\sim}{B}_{EV}$ is given by

$$\underset{\sim}{b}_{EV} = (\underset{\sim}{X}_n^T \underset{\sim}{W}_n \underset{\sim}{X}_n - n \underset{\sim}{\Sigma}_{uu})^{-} \underset{\sim}{X}_n^T \underset{\sim}{W}_n \underset{\sim}{Y}_n \tag{2.8}$$

where the elements of $\underset{\sim}{X}_n^T \underset{\sim}{W}_n \underset{\sim}{X}_n$ and of $\underset{\sim}{X}_m^T \underset{\sim}{W}_n \underset{\sim}{Y}_n$ are defined as previously provided. A consistent estimator for the covariance matrix of $\underset{\sim}{b}_{EV}$ is given by

$$\hat{\underset{\sim}{V}}_{EV} = (\underset{\sim}{X}_n^T \underset{\sim}{W}_n \underset{\sim}{X}_n - n \underset{\sim}{\Sigma}_{uu})^{-} \hat{\underset{\sim}{G}}_{EV} (\underset{\sim}{X}_n^T \underset{\sim}{W}_n \underset{\sim}{X}_n - n \underset{\sim}{\Sigma}_{uu})^{-}$$

with the elements of $\hat{\underset{\sim}{G}}_{EV}$ defined as in (2.5) with

$$\hat{d}_{hjkr} = X_{hjkr} \hat{v}_{hjk} w_{hjk} ,$$

$$\hat{v}_{hjk} = Y_{hjk} - \sum_{r=1}^{p} b_{EV,r} X_{hjkr} .$$

Regularity conditions for the above covariance matrix's consistency have also been provided by Fuller (1975).

3. MONTE-CARLO STUDY USING CPS DATA

The data used for this investigation were those used by Frankel [1971] and were collected by the U.S. Bureau of the Census in the March 1967 Current Population Survey [1963]. The finite population

consisted of 45,737 observations grouped in 3240 primary units. Two sample designs were used in this investigation. In sample design I the original 3,240 primary units in the population were divided into 6 strata containing 540 primary units each. In sample design II, the 3,240 primary units were divided into 12 strata, each of size 270 primary units. This stratification was carried out by splitting each of the 6 strata used in design I into two strata. In sample designs I and II, two primary sampling units were selected s.r.s. without replacement from each stratum of the population. The data was stored on a tape. Each individual element stored on this tape was identified by a household number and a p.s.u. code. The p.s.u. numbers were ordered from 1 to 3,240. All the elements associated with a specific p.s.u. were grouped together within the strata defined by the position of the p.s.u. on the sequence. In the case of the 6 strata design, the first 540 p.s.u. made up stratum I, the second 540 p.s.u. made up stratum II, etc. In the case of the 12 strata design, each stratum was arranged in a sequence of 270 p.s.u. Each of the two sampling designs called for the selection of two primary sampling units from each stratum of the population. A computer program was written to select the two primary sampling units using a simple random without replacement sampling scheme. For sample design I, 6 independent pairs of random numbers were generated. Each element of the pair was generated by a uniform (0,1) random number generator. The elements of each pair were multiplied by 540 and the product was truncated. For sample design II, 12 independent pairs of random numbers were generated, with each element of the pair generated by a uniform (0,1) random number generator. The elements of each pair were multiplied by 270. Two hundred independent samples were selected in this manner for each sampling design.

The dependent variable of interest was log Income of the household head and the independent variables were age, age squared and education. To ensure that the matrix of sums of squares and products of the independent variables was nonsingular, the independent variables were coded as: $\text{Age} - 43$, $(\text{Age} - 43)^2 - 70$ and $\text{Education} - 12$.

Let X_{hjkr} $(r = 1, 2, \ldots, 4)$ denote the value of the r-th independent variable and Y_{hjk} the value of the dependent variable for the k-th element $(k = 1, 2, \ldots, M_{ij})$ in the j-th p.s.u. $(j = 1, 2, \ldots, N_i)$ of the h-th stratum $(h = 1, 2, \ldots, L)$ of the population. Similarly, x_{hjkr} denotes the value of the r-th independent $(r = 1, 2, \ldots, 4)$ variable and Y_{hjk} the value of the dependent variable for the k-th element $(k = 1, 2, \ldots, M_{ij})$ in the j-th primary $(j = 1, 2)$ of the h-th stratum $(h = 1, 2, \ldots, L)$ of a selected sample. In addition, define

$$X_{hjk1} = 1 \quad \text{and} \quad x_{hjk1} = 1$$

for all h, j and k. The sampling behavior of the two sets of statistics associated with the O.L.S. estimator given by (2.3) and those associated with the error-in-variables procedure given by (2.8) will be investigated.

For the case of errors-in-variables, it is assumed that the response errors are independent between secondary units (clusters in our case) within the same primary unit (stratum) as well as between secondary units on different primary units. For the errors-in-variables model, age and education were observed subject to response error. Using the U.S. Bureau of the Census (1972) coding study, response variances, for Age - 43, $(\text{Age} - 43)^2 - 70$ and Education - 12, were assumed to be 0.3, 91.0 and 3.0 respectively. It was assumed that the response error of variance was uncorrelated with that of age and education. In our case,

$$\underset{\sim}{\Sigma}_{uu} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & \\ 0 & 0 & 91 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

The "t-statistics" are given by

$$t(b_r) = \frac{b_r - B_r}{s(b_r)} \quad r = 1, 2, \ldots, 4$$

where b_r, B_r and $s(b_r)$ are the sample regression estimates, population

regression parameters and standard deviation of b_r respectively for the O.L.S. or E.V. case. The properties of t-statistics are also of interest.

4. RESULTS FROM THE MONTE-CARLO STUDY

The data obtained in the 200 samples for each sample design was used for both regression procedures. The results of the two experiments are presented in several tables. Table 1 gives for each experiment the mean and variance for the regression coefficients. Note that the standard error in design I are approximately $\sqrt{2}$ times the standard errors of the corresponding coefficient in design II. This is to be expected, since the number of primary sampling units in the 12 strata design is twice the number in the 6 strata design. From Table 2, considering the ratio of the estimated bias of 200 sample regression estimates to the estimated standard error of their mean to be distributed as Student's t with 199 d.f., we conclude that the bias is reduced as the sample size increases.

Additional information concerning the frequency distributions of the estimates computed in the Monte-Carlo study is given in Tables 3 and 4 which contains the observed percentiles of the calculated t's. Examination of Tables 3 and 4 reveals that the distribution for $t(b)$ and $t\left(b(e)\right)$ agrees more closely with the theoretical t distribution near the median than in the tails. Comparisons of the 1%, 5%, 95% and 99% points for the t statistics in Tables 3 and 4 reveal the effects of increased sample size. For instance, in Table 3 the 5% and 95% points for $t\left(b_3(e)\right)$ are -2.321 and 2.573 which are considerably higher than the corresponding points for the t distribution with 6 degrees of freedon, ± 1.943. For these same statistics, the 5% and 95% points in Table 4 are -1.641 and 2.076 as compared to ± 1.782, the corresponding points for the t distribution with 12 degrees of freedom. These observations suggest that the variances of the sample regression coefficients estimates have been underestimated in small samples, though not by much.

Comparing the results for O.L.S. regression coefficients in Table 5, it is evident that Frankel's calculated t's for these coefficients are closer to the theoretical t distribution than the ones obtained in our study. One explanation for this is that only urban males between the ages 28-58 were selected for our study. This resulted in decreasing the average number of elements in the sample for designs I and II from 170.3 and 339.5 (as used for Frankel's study) to 61.5 and 124.5 (as used for our study) respectively.

In summary, the results of this investigation indicate the sample estimates of the multiple regression coefficients have small biases, and the distribution of the t-statistics computed for both the O.L.S. and errors-in-variables procedures are well approximated by the theoretical t distribution. In addition, the agreement improves as the number of strata used in the design increases.

TABLE 1: Means of 200 Regression Sample Vectors

Number of strata in experiment	Regression coefficients	LEAST-SQUARES MODEL				ERRORS-IN-VARIABLES MODEL			
		b_1	b_2	b_3	b_4	$b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
6	Population value	8.9289	0.0029	-0.0007	0.0846	8.9405	0.0053	-0.0006	0.1194
6	Means of 200 samples	8.9115	0.0027	-0.0009	0.0812	8.9207	0.0055	-0.0008	0.1213
6	Estimated Standard deviation of estimates	0.1136	0.0112	0.0013	0.0308	0.1082	0.0116	0.0015	0.0477
6	Estimated standard error of mean	0.0080	0.0008	0.0001	0.0022	0.0076	0.0008	0.0001	0.0034
12	Population value	8.9289	0.0029	-0.0007	0.0846	8.9405	0.0053	-0.0006	0.1194
12	Mean of 200 samples	8.9254	0.0039	-0.0006	0.0842	8.9344	0.0068	-0.0006	0.1225
12	Estimated standard deviation of estimates	0.0724	0.0075	0.0008	0.0228	0.0712	0.0078	0.0009	0.0332
12	Estimated standard error of mean	0.0051	0.0005	0.0001	0.0016	0.0050	0.0005	0.0001	0.0023

TABLE 2: Estimated Bias of Regression Estimates for 200 Replicates

Number of strata in experiment	LEAST-SQUARES MODEL				ERRORS-IN-VARIABLES MODEL			
	b_1	b_2	b_3	b_4	$b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
6	-0.0174*	-0.0002	-0.0002*	-0.0034	-0.0198*	0.0002	0.0002	0.0019
12	-0.0035	0.0010	0.0001	-0.0004	-0.0061	-0.0015*	0.0000	-0.0031

*significant at the 5% level

TABLE 3: Comparison of Observed Percentiles of the Calculated t's with the theoretical percentiles for the t distribution with 6 degrees of freedom

Probability in percent	Theoretical percentile for student's t	OBSERVED PERCENTILE for t(b)				OBSERVED PERCENTILE for t(b(e))			
		b_1	b_2	b_3	b_4	$b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
1	-3.143	-4.841	-3.794	-3.479	-5.315	-4.720	-3.316	-3.329	-5.545
5	-1.943	-2.616	-2.053	-2.278	-2.650	-2.366	-2.055	-2.321	-2.203
10	-1.440	-1.855	-1.734	-1.652	-1.773	-1.667	-1.547	-1.615	-1.811
20	-0.906	-1.070	-1.131	-1.153	-1.314	-0.973	-0.908	-1.110	-1.082
30	-0.553	-0.695	-0.625	-0.731	-0.823	-0.623	-0.515	-0.825	-0.637
40	-0.265	-0.392	-0.258	-0.481	-0.434	-0.328	-0.263	-0.535	-0.395
50	-0.000	-0.057	-0.037	-0.211	-0.222	-0.053	0.016	-0.202	-0.102
60	0.265	0.221	0.298	0.153	0.083	0.212	0.262	0.126	0.213
70	0.553	0.630	0.632	0.478	0.468	0.428	0.677	0.459	0.551
80	0.906	1.236	0.902	0.906	0.982	1.104	0.932	0.826	0.845
90	1.440	1.828	1.736	1.874	1.664	1.774	1.564	1.799	1.450
95	1.943	2.567	2.898	2.789	2.148	2.849	2.142	2.573	1.666
99	3.143	4.418	4.679	5.116	3.638	5.022	3.852	4.830	3.351

TABLE 4: Comparison of Observed Percentiles of the Calculated t's with the theoretical percentiles for the t distribution with 12 degrees of freedom

Probability in percent	Theoretical percentile for student's t	OBSERVED PERCENTILE t(b)				OBSERVED PERCENTILE $t(b_e)$			
		b_1	b_2	b_3	b_4	$b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
1	-2.681	-2.442	-3.167	-2.545	-3.278	-2.694	-2.679	-2.526	-3.222
5	-1.782	-1.777	-1.813	-1.536	-2.306	-1.822	-1.659	-1.641	-1.659
10	-1.356	-1.364	-1.294	-1.316	-1.440	-1.258	-1.273	-1.308	-1.236
20	-0.873	-0.975	-0.666	-1.004	-0.961	-0.842	-0.693	-0.863	-0.796
30	-0.539	-0.554	-0.364	-0.623	-0.451	-0.511	-0.407	-0.542	-0.309
40	-0.253	-0.195	-0.124	-0.301	-0.145	-0.298	-0.090	-0.161	-0.036
50	0.000	0.076	0.138	0.032	0.056	-0.024	-0.326	0.122	0.195
60	0.253	0.277	0.492	0.378	0.306	0.266	0.554	0.412	0.379
70	0.539	0.640	0.756	0.666	0.694	0.504	0.784	0.653	0.636
80	0.873	0.981	1.141	1.043	0.987	0.938	1.114	0.963	0.930
90	1.356	1.543	1.709	1.644	1.538	1.566	1.735	1.516	1.345
95	1.782	1.976	2.016	1.951	2.028	1.848	2.112	2.076	1.757
99	2.681	2.860	2.786	3.300	2.944	2.824	3.057	3.284	2.428

TABLE 5: Comparison of Observed Proportion for Calculated t(b) within stated limits to the theoretical proportion for the t distribution

NUMBER OF STRATA IN EXPERIMENT	INTERVALS	THEORETICAL PROPORTION	OBSERVED PROPORTION	
			FRANKEL'S STUDY	OUR STUDY
6	± 2.576	0.9580	0.9421	0.9350
6	± 1.960	0.9023	0.8733	0.8525
6	± 1.645	0.8489	0.8146	0.8104
6	± 1.282	0.7529	0.7167	0.7037
6	± 1.000	0.6441	0.6029	0.5950
12	± 2.576	0.9757	0.9662	0.9640
12	± 1.960	0.9264	0.9121	0.9103
12	± 1.645	0.8741	0.8496	0.8447
12	± 1.282	0.7760	0.7437	0.7500
12	± 1.000	0.6630	0.6217	0.6100

REFERENCES

Frankel, M.R. (1971), Inference from Survey Samples. Published Ph.D. dissertation. University of Michigan, Ann Arbor, Michigan.

Fuller, W.A. (1975), Regression Analysis For Sample Survey. Sankhyā, 37, Series C, pg. 117-132.

Fuller, W.A. (1981), Properties of some estimators for the errors-in-variables model. Annals of Statistics, 8, pg. 407-422.

Hartley, H.O. and Sielken, R.L. (1975), A "super-population viewpoint" for finite population sampling. Biometrics, 31, pg. 411-422.

Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1980). SUPER CARP: Sixth Edition, Statistical Laboratory Survey Section, Iowa State University, Ames, Iowa.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980), Regression of Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, Ser. A, 143, pg. 474-487.

Konijn, H.S. (1962), Regression Analysis in Sample Surveys. Journal of the American Statistical Association 57, pg. 590-606.

Särndal, C.E. (1978), Design-based and model-based inference in survey sampling. Scand. J. Statist., 5, pg. 27-52.

Scott, A.J., and Holt D. (1982), The Effect of two-stage sampling on ordinary least square methods. J.A.S.A. 77, pg. 848-854.

Shah, B.V., Holt, H.M., and Folsom, R.E. (1977), Inference about regression models from sample survey data. Int. Assoc. of Survey Statisticians, 3rd. Meeting, New Delhi.

Tepping, B.J. (1968), The Estimation of Variance in Complex Surveys, Proc. Inc. Statist. Sec., Amer. Statist. Assoc., 66, pg. 411-414.

U.S. Bureau of the Census (1972). Evaluation and Research Program of the U.S. Census of Population and Housing 1960: Effects of different reinterview techniques on estimates of simple response variance. U.S. Bureau of the Census, Series ER60, No. 11.