Journal of Official Statistics
Vol. 2. No. 1, 1986, pp. 3–11

Estimation of Regression Parameters for Finite Populations: A Monte-Carlo Study

Michael A. Hidiroglou[1]

Abstract: Regression analysis is quite often performed using data arising from complex survey designs. These designs will invariably be stratified and clustered with at least one stage of selection. Inferences from such regression analyses must be based on properly computed variance-covariance matrices of the regression coefficients taking the survey design into account. A Monte-Carlo study was conducted to study the small sample properties of the ordinary least squares regression estimator and an errors-in-variables regression estimator, given that the sample design had been taken into account. The elements of the sample regression vectors were normalized by subtracting the corresponding elements of the population regression vector and dividing the difference by the estimated standard error. The distribution of the resulting statistics, termed "t-statistics," was investigated. It was determined that the distribution agreed with that of Student's t. The sample regression vector was approximately unbiased for the population regression vector.

Key words: Primary sampling unit; complex survey design; regression; standard error; simulation; finite populations.

1. Introduction

Regression analysis is a 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 often ignore the assumptions that should be supported by the data sets they analyse. Regression analysis procedures in standard computer packages require that the observations are independent. This crucial assumption is violated if the data has been collected using cluster or multistage sampling designs. Subsequent analyses using these packages do not take this important consideration into account, with the net effect being that the standard estimators for the variance of the estimated regression coefficients are likely to be seriously underestimated. Test statistics and confidence regions based on those variance estimators are consequently 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 et al. (1980), Särndal (1978), and by Scott and Holt (1982). These authors have pointed out the dangers of using traditional

[1] Statistics Canada. The author would like to thank the associate editor and two referees for their constructive comments.

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), Kish and Frankel (1974), 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. Frankel used the Taylor approximations to the variance formula, suggested by Tepping (1968) to obtain variance estimates of the regression coefficients, in order to compute the variance-covariance matrix of the estimated sample regression coefficients. The data used for his 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 his regression analysis was the estimation of the finite population parameters as defined by the population moments.

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. These small sample properties will be determined using a simulation carried out on the data from the 1967 Current Population Survey. Two types of regression procedures will be studied. One, where the data are not subject to measurement error and the other, where the data are 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.

The simulation was carried out using the computer program SUPER CARP (see Hidiroglou et al. (1980)). 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

$$\mathbf{y}_N = \mathbf{x}_N \mathbf{B}_{\text{OLS}} + \mathbf{e}_N, \tag{2.1}$$

where $\mathbf{y}_N$ is an N x 1 vector of observations on the dependent variable; $\mathbf{x}_N$ is an N x p non-stochastic matrix of observations on p independent variables; $\mathbf{B}_{\text{OLS}}$ is the p-dimensional vector of regression coefficients; $\mathbf{e}_N$ represents an N x 1 vector of deviations from the linear relationship. N is the size of the population of interest. In the absence of measurement error on $\mathbf{y}$ and $\mathbf{x}$, minimizing the sum of the squared deviations over the entire population yields the following definition (Frankel (1971)) for the population regression coefficients

$$\mathbf{B}_{\text{OLS}} = (\mathbf{x}_N^T \mathbf{x}_N)^{-} \mathbf{x}_N^T \mathbf{y}_N, \tag{2.2}$$

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

In this study, the estimator for $\mathbf{B}_{\text{OLS}}$ is obtained via a one-stage stratified clustered sample obtained as follows. The population is first divided into L strata (labelled $h = 1, 2, \ldots, L$). For each stratum, a sample of size n_h is drawn without replacement from N_h

clusters and from each selected cluster of size M_{hj}, all the elements are selected. A natural estimator for $\mathbf{B}_{\text{OLS}}$ which appeals to survey statisticians is one that takes the survey weights into account. A survey weight is defined as the inverse of the probability of inclusion of an ultimate unit.

The sample estimator for $\mathbf{B}_{\text{OLS}}$ is given by

$$\mathbf{b}_{\text{OLS}} = (\mathbf{x}_n^T \mathbf{W}_n \mathbf{x}_n)^- \mathbf{x}_n^T \mathbf{W}_n \mathbf{y}_n, \quad (2.3)$$

where $\mathbf{x}_n$ is an n x p matrix of observed independent variables, $\mathbf{y}_n$ is an n x 1 vector of observed dependent variables in the sample, and $\mathbf{W}_n$ is an n x n diagonal matrix consisting of the survey weights. The rs-th element of $\mathbf{x}_n^T \mathbf{W}_n \mathbf{x}_n$ 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 variable,

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

Similarly, the s-th element of $\mathbf{x}_n^T \mathbf{W}_n \mathbf{y}_n$ 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 $\mathbf{b}_{\text{OLS}}$ under certain regularity conditions to ensure the convergence to normality and consistency. This estimator is given by

$$\hat{\mathbf{V}}_{\text{OLS}} = (\mathbf{x}_n^T \mathbf{W}_n \mathbf{x}_n)^- \hat{\mathbf{G}}_{\text{OLS}} (\mathbf{x}_n^T \mathbf{W}_n \mathbf{x}_n)^-, \quad (2.4)$$

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

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

with

$$f_h = n_h / N_h,$$

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

$$\hat{v}_{hjk} = y_{hjk} - \sum_{r=1}^{p} \mathbf{b}_{\text{OLS}}(r) x_{hjkr},$$

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

and

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

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

$$\mathbf{Y}_N = \mathbf{X}_N \mathbf{B}_{\text{EV}} + \mathbf{e}_N, \quad (2.6)$$

where $\mathbf{Y}_N$ and $\mathbf{X}_N$ are the observed random variables incorporating measurement error. That is

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

where $\underset{\sim}{\delta}_N = (\underset{\sim}{\varepsilon}_N, \mathbf{u}_N)$ is the matrix of response errors for the population. The following assumptions are made concerning $\underset{\sim}{\varepsilon}_N$ and $\mathbf{u}_N$:

i) $E\{\underset{\sim}{\delta}_{hjk} | (y_{hjk}, \mathbf{x}_{hjk})\} = \mathbf{0}$ for all vectors

$(y_{hjk}, \mathbf{x}_{hjk})$; $h = 1, \ldots, L; j = 1, \ldots, N_h;$

$k = 1, \ldots M_{hj};$

where $\underset{\sim}{\delta}_{hjk} = (\varepsilon_{hjk}, \mathbf{u}_{hjk})$ is the hjk-th row vector of $\underset{\sim}{\delta}_N$.

ii) $\delta_{h'j'k'}$ is independent of δ_{hjk} for $h \neq h'$ or $j \neq j'$or $k \neq k'$. That is, the measurement errors on different units are independent.

iii) The covariance matrix of $\underset{\sim}{\delta}_{hjk}$ is given by

$$\underset{\sim}{\Sigma}_{\delta} = \begin{pmatrix} \sigma_{\varepsilon}^2, & \underset{\sim}{\Sigma}_{\varepsilon u} \\ \underset{\sim}{\Sigma}_{u\varepsilon}, & \underset{\sim}{\Sigma}_{uu} \end{pmatrix},$$

and it is assumed to be known.

Similarly to the ordinary least squares case ($\mathbf{B}_{\text{OLS}}$), $\mathbf{B}_{\text{EV}}$ may be defined as

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

A sample estimator for $\mathbf{B}_{\text{EV}}$ is given by

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

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

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

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

$$\hat{d}_{hjkr} = X_{hjkr} \hat{v}_{hjk} - \hat{v}_{hjk}^2 \left[\frac{\sigma_{u_r\varepsilon} - \sum_{s=1}^{p} \sigma_{u_r u_s} b_{\text{EV}}(s)}{\hat{\sigma}_v^2} \right] w_{hjk},$$

$$\hat{v}_{hjk} = Y_{hjk} - \sum_{s=1}^{p} b_{\text{EV}}(s) X_{hjks},$$

$$\hat{\sigma}_v^2 = \sigma_\varepsilon^2 - 2 \underset{\sim}{\Sigma}_{\varepsilon u} \mathbf{b}_{\text{EV}} + \mathbf{b}_{\text{EV}}^T \underset{\sim}{\Sigma}_{uu} \mathbf{b}_{\text{EV}}.$$

Regularity conditions for the above covariance matrix's consistency have also been provided by Fuller (1975). Extensions to the case where Σ_δ has been estimated have been provided by Fuller (1975), Fuller and Hidiroglou (1978), and Fuller (1981). The required computational forms of these extensions are described in Hidiroglou et al. (1980).

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. The finite population consisted of 45 737 households grouped in 3 240 primary sampling 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 stratum having 270 primary units. This stratification was carried out by splitting each of the six strata used in design I into two strata. In sample designs I and II, two primary sampling units (p.s.u.) were selected by s.r.s. without replacement from each stratum of the population. The data were 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 six-strata design, the first 540 p.s.u.'s made up stratum 1, the second 540 p.s.u.'s made up stratum 2, etc. In the case of the twelve-strata design, each stratum was arranged in a sequence of 270 p.s.u.'s. Only urban males in the ages 28–58 were selected for this study. 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, six 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. Each pair of random numbers was used to select the two p.s.u.'s within each of the strata. All the elements (M_{hj}) within a selected p.s.u. (j) and stratum (h) were included in each of the samples. The sample design for this experiment may therefore be described as a one stage stratified cluster design with the clusters being selected using simple random sampling without replacement. Two hundred independent samples were selected in this manner for each sampling design.

The dependent variable in the regression was log income of the household head and the independent variables included an intercept term (variable 1), age (variable 2), age squared (variable 3) and education (variable 4). To ensure that the matrix of sums of squares and products of the independent variables was nonsingular, the last three independent variables (2–4) were coded as: Age–43, $(\text{Age}-43)^2 - 70$ and Education – 12.

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 errors-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 assumed to be 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 log income was uncorrelated with that of age and education. In our case,

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

The t-statistics are given by

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

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. Means of 200 Regression Sample Vectors

Number of Strata in Experiment	Regression Coefficients	Least-Squares Model b_1	b_2	b_3	b_4	Errors-in-Variables Model $b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
6	Population value	8.9289	0.0029	−0.0007	0.0846	8.9405	0.0053	−0.0006	0.1194
6	Observed mean of 200 regression coefficient estimates	8.9115	0.0027	−0.0009	0.0812	8.9207	0.0055	−0.0008	0.1213
6	Observed standard deviation of 200 regression coefficient estimates	0.1136	0.0112	0.0013	0.0308	0.1082	0.0116	0.0015	0.0477
6	Observed mean of 200 standard error estimates	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	Observed mean of 200 regression coefficient estimates	8.9254	0.0039	−0.0006	0.0842	8.9344	0.0068	−0.0006	0.1225
12	Observed standard deviation of 200 regression coefficient estimates	0.0724	0.0075	0.0008	0.0228	0.0712	0.0078	0.0009	0.0332
12	Observed mean of 200 standard error estimates	0.0051	0.0005	0.0001	0.0016	0.0050	0.0005	0.0001	0.0023

Table 1 gives, for each experiment, the mean and variance for the regression coefficients. Note that the standard errors 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 six-strata design.

Table 2. Estimated Bias of Regression Estimates for 200 Replicates

Number of Strata in Experiment	Least-Squares Model b_1	b_2	b_3	b_4	Errors-in-Variables Model $b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
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.

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 contain the observed percentiles of the calculated t's. Examination of Tables 3 and 4 reveals that the distribution

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 $t(b)$ b_1	b_2	b_3	b_4	Observed Percentile for $t[b(e)]$ $b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
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	4.890	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 $t(b)$ b_1	b_2	b_3	b_4	Observed Percentile $t(b_e)$ $b_1(e)$	$b_2(e)$	$b_3(e)$	$b_4(e)$
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

for $t(b)$ and $t[b(e)]$ 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 $t[b_3(e)]$ are -2.321 and 2.573 which are considerably higher than the corresponding points for the t distribution with 6 degrees of freedom, ± 1.943. The degrees of freedom can be observed from equations (2.5) as being equal to the total number of selected p.s.u.'s minus the total number of strata. 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.

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

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 that 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.

5. 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, Series C, 37, pp. 117–132.

Fuller, W. A. and Hidiroglou, M. A. (1978): Regression Estimation after Correcting for Attenuation. Journal of the American Statistical Association, 73, pp. 99–104.

Fuller, W. A. (1981): Properties of Some Estimators for the Errors-in-Variables Model. Annals of Statistics, 8, pp. 407–422.

Hartley, H. O. and Sielken, R. L. (1975): A "Super-Population Viewpoint" for Finite Population Sampling. Biometrics, 31, pp. 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 Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, Series A 143, pp. 474–487.

Kish, L. and Frankel, M.R. (1974): Inference from Complex Samples. Journal of the Royal Statistical Society, Series B, 36, pp. 1–37.

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

Särndal, C. E. (1978): Design-Based and Model-Based Inference in Survey Sampling. Scandinavian Journal of Statistics, 5, pp. 27–52.

Scott, A. J. and Holt, D. (1982): The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. Journal of the American Statistical Association, 77, pp. 848–854.

Shah, B. V., Holt, H. M., and Folsom, R. E. (1977): Inference about Regression Models from Sample Survey Data. Bulletin of the International Statistical Institute, 47, Book 3, pp. 43–57.

Tepping, B. J. (1968): The Estimation of Variance in Complex Surveys. American Statistical Association, Proceedings of the Social Statistics Section, pp. 11–18.

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.

Received November 1984
Revised October 1985