SUPER CARP

Michael A. Hidiroglou, Statistics Canada
Wayne A. Fuller, Iowa State University
Roy D. Hickman, Iowa State University

SUMMARY

SUPER CARP is a program designed for several types of analyses of survey data. The program can be used to construct estimated totals, ratio estimates, the difference of ratio estimates, and regression estimates for multistage stratified samples. It contains a number of procedures appropriate for data observed subject to response (measurement) error. The program can also be used to compute the coefficients of regression equations and the standard errors of the regression coefficients for such designs. The variances of the estimates are estimated using the Taylor series approximation. Four regression options are available, including three types of estimation for data subject to response error.

1. INTRODUCTION

In many scientific investigations in the social sciences, statistics are computed for data collected in large complex surveys. The statistics computed from such surveys range from simple statistics such as totals and means to more complex statistics such as regression coefficients. SUPER CARP is a program package designed to compute such statistics and the estimated variances of the statistics. The variances of the statistics are computed using Taylor methods. Some of the theory underlying the techniques is given in Fuller (1971), Fuller (1975), and Fuller and Hidiroglou (1978). The program is written in double precision FORTRAN G, and was developed on an IBM 360/370 system.

2. GENERAL DESCRIPTION

The data are read into the computer identified by stratum and cluster (primary sampling unit). Simpler sample designs may be read with less identification. In the general form, a q-dimensional data vector is read for each observation. We denote this data vector by

$$(Z_{ijk1}, Z_{ijk2}, \cdots, Z_{ijkq}),$$

where $i = 1,2,\ldots,L$ denotes strata, $j = 1,2,\ldots,n_i$ denotes the number of sampled clusters within the i^{th} stratum, and $k = 1,2,\ldots,m_{ij}$ denotes the number of sampled elements with the j^{th} cluster within the i^{th} stratum. Z_{ijkr} is the ijk^{th} observation for the r^{th} variable, $r = 1,2,\ldots,q$. Weights associated with the ijk^{th} observation are denoted by w_{ijk}. The weights are usually inversely proportional to the selection probabilities. For sample designs where the selection process is without replacement, the user may enter sampling rates to be used in constructing finite population correction factors for each stratum. The user selects a type of analysis and p variables out of the q variables to enter the analysis.

3. INPUT AND STORAGE

The data that are read into the program must have proper identification. In a typical survey situation, a stratum identification, a cluster (primary sampling unit) identification and a weight are required. The observations must be ordered by clusters within strata. Hence, the ordered data sequence defines the stratum and cluster structure of the data. The standard input is in the form of punched cards, but data may be read from tape or disk. The maximum number of variables that can be read into the standard program is 50, while the maximum number of variables that can be included in one analysis is 20.

Program instructions are input on control cards. There are six mandatory control cards; the parameter card, the format card, the screening card, the finite population correction card, the analysis card, and the variable identification card. The parameter card contains information such as the problem name, the number of observations in the sample, the number of variables to be input, the type of survey data, and the number of analyses to be performed.

The format card specifies the input format for the data. The screening card specifies operations which reject missing observations or define tolerance limits for the input data. The finite population correction card is used to input sampling rates to be used in constructing finite population correction factors. The analysis card specifies the type of analysis (e.g. regression, ratio estimation, total estimation). Also included in the analysis card is the number of variables to be included in the analysis and other characteristics of certain analyses. The variable identification card identifies the variables to be used in the chosen analysis. Optional cards are used to identify the screening operations, and to input error variances for the errors in variables analysis. The program can process several sets of data at one time.

Each input data vector is augmented by an element identically equal to 1 when an intercept is to be included in any of the regression analyses. The resulting data vector is put through any required screening and then stored in temporary disk storage. The data set on disk storage is input into a subroutine which will collapse any stratum which has only one element (this collapsing is done with respect to adjacent strata). New sampling rates will be computed if collapsing is necessary and if sampling rates have been originally input.

4. COMPUTATIONS

A recursive method for computing means, corrected sums of squares and cross products has been

adopted based upon the algorithm given by Neely (1966). This method provides more accuracy than the usual 'desk-machine' method and also requires only one computer pass of the data set.

The algorithm developed by Healy (1968) is used for matrix inversion. Using this algorithm, positive semi-definite matrices are inverted. If the original matrix is not of full rank, the resulting inverse is a generalized inverse in which the row and column corresponding to the redundant parameter are zero. This inversion algorithm operates by using the Cholesky decomposition.

The root computation associated with errors in variables requires the smallest root of the determinantal equation

$$|\underset{\sim}{A} - \lambda \underset{\sim}{B}| = 0 ,$$

where $\underset{\sim}{A}$ is positive definite symmetric. The matrix $\underset{\sim}{A}$ can be decomposed into the product $\underset{\sim}{U}\,\underset{\sim}{U}'$, where $\underset{\sim}{U}$ is the Cholesky decomposition of $\underset{\sim}{A}$. The determinantal equation can then be written

$$|\underset{\sim}{I} - \lambda \underset{\sim}{C}| = 0 ,$$

where $\underset{\sim}{C} = \underset{\sim}{U}^{-1}\,\underset{\sim}{B}\,\underset{\sim}{U}^{-1'}$. The eigenvalues are obtained using the diagonalization method originated by Jacobi and adapted by von Neumann for large computers. The method is described in Ralston and Wilf (1962, Ch. 7).

The program can be used to compute tests of hypothesis for any subset of the regression parameters. Hence if $\underset{\sim}{b} = (b_1, b_2, \ldots, b_p)$ estimates $\underset{\sim}{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, then hypotheses of the form

$$H_o: \beta_{i1} = \beta_{i2} = \cdots = \beta_{ih} = 0, \quad (h \leq p)$$

against the alternative

$$H_1: \beta_{ij} \neq 0 \text{ for some } ij ,$$

may be tested. This test statistic is approximately distributed as Snedecor's F under the null hypothesis.

5. OUTPUT

The printed output includes the problem identifier specified on the parameter card, the total sample size, the total number of input variables, the intercept indicator, the type of data identification, the number of analyses, and a listing of the data, if requested. The analyses identified by number and type follow. For the regression analyses, means, variances, and covariances of the dependent and independent variables are given. The regression coefficients are listed with their associated standard errors and 't-statistics.'

A copy of the SUPER CARP manual, containing instructions for program use and a description of computations, will be sent upon request. A copy of the program on tape can be obtained for $25 by writing the Survey Section, Department of Statistics, Iowa State University, Ames, Iowa 50011.

REFERENCES

Fuller, W. A. (1971), Properties of estimators in the errors-in-variables model, presented at the 1971 annual meeting of the Econometric Society.

Fuller, W. A. (1975), Regression analysis for sample survey. Sankhyā C 37, 117-132.

Fuller, W. A., and Hidiroglou, M. A. (1978), Regression estimation after correcting for attenuation. Journal of the American Statistical Association 73, 99-104.

Healy, M. J. R. (1968), Triangular decomposition of a symmetric matrix. Applied Statistics 17, 195-197.

Healy, M. J. R. (1968), Inversion of a positive semi-definite symmetric matrix. Applied Statistics 17, 198-199.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1978), SUPER CARP, Survey Section, Iowa State University, Ames, Iowa.

Neely, P. M. (1966), Comparison of several algorithms for computation of means, standard deviations and correlation coefficients. Comm. A.C.M. 9, 496-499.

Ralston, A., and Wilf, H. S. (Eds.)(1962), Mathematical Methods for Digital Computers. Wiley, New York.