RECENT DEVELOPMENTS IN PC CARP

William J. Kennedy, Ouhong Wang and Wayne A. Fuller
Wayne A. Fuller, Iowa State University, Department of Statistics
Ames, IA 50011

KEY WORDS: Variance estimation, complex surveys, computer program

Extensions of the program PC CARP to batch processing on UNIX workstations is described. The use of the basic algorithms of PC CARP as a part of a processing program for a large nationwide survey conducted by the Soil Conservation Service is discussed.

1. INTRODUCTION

The computer program PC CARP was developed at Iowa State University under an agreement with the International Statistical Programs Center of the U.S. Bureau of the Census. The first version was released in December 1986. The program was designed to compute estimates and estimated variances for quantities common to survey sampling. Estimated totals, means, ratios and proportions and their estimated standard errors can be computed for subpopulations. The subpopulation specification is very flexible permitting subpopulations to be defined by high order cross classifications. The program can also be used to compute regression equations, two—way tables and quantiles.

The program contains an option to compute variances for two—stage samples. The program will collapse one—unit strata by adjoining the single unit to the adjacent stratum. PC CARP was heavily based on the algorithms of SUPER CARP, described in Hidiroglou, Fuller and Hickman (1976), and on the regression procedures of Hidiroglou (1974). With the exception of Quantiles, the Taylor method of variance computation is used. See, for example, Binder (1983) and Fuller (1975). Quantiles can be computed for subpopulations. Confidence intervals for quantiles are computed by inverting confidence intervals for the cumulative distribution function. The basic theory for quantiles is discussed by Francisco and Fuller (1991).

PC CARP requires a complete data set, but a simple imputation program is provided as a means of removing missing values.

The original program was created for use in developing countries on the IBM Personal Computer XT. The program can handle an unlimited number of observations in an unlimited number of strata and requires 410K bytes of memory and a math coprocessor.

Because of the original objectives, it was important that the program be relatively easy to use by those with minimal training in survey sampling. Therefore, PC CARP is menu driven with "help" and "go—back" capabilities. The program has been well received and several hundred copies have been distributed throughout the world. Plewis (1989) is a review of the program.

Two supplements to PC CARP were produced after the initial release. The first, authored by Jorge Morel and released in January 1988, computes estimates of the parameters of the multinomial logistic function for designs of the type covered by PC CARP. See Morel (1989). The second, authored by Gary Sullivan and released in August, 1988, computes the variance of estimators constructed from surveys that have been post stratified. There have been no additional releases and little modification of the program since 1988.

2. UNIX VERSION OF PC CARP

PC CARP was originally designed to run on IBM compatible personal computers to process unlimited numbers of observations, with moderately flexible output options. The user interface routines, which support interactive problem and analysis specifications, were written in assembly language.

Experience in the use of this software proved its utility. It also showed the need for a batch processing capability and versions of the program which will execute in other environments.

In order to assess the difficulty of transporting the program to a distributed computing environment, a preliminary development project was carried out at Iowa State University. The objective was to implement PC CARP, without major modification, on the

ULTRIX—based distributed network using DECstation 3100 and 5000 machines. The user interface was rewritten using the CURSES support library routines to allow cursor control. Needed modifications inside the numerical computing routines were made so that the workstation version would execute correctly. A crude but effective batch capability was introduced, primarily to save time in problem definition. The results of this project were encouraging. Numerical results, approaching the accuracy of those obtained in the PC, were obtained in the workstation version. An excellent product in this environment seemed entirely possible without completely reworking the numerical routines which would have been a huge task.

Work is now under way to extend the workstation version in several directions. The user interface is being redesigned and programmed to better employ the powerful X Window System support. Background processes will be produced from user interactive problem and analysis specifications to yield a desirable form of batch processing. Problem size will be extended to 500 variables per problem from the current 50 variable maximum in the PC version. These modifications will produce a program having essentially the same analysis capabilities as the PC version, but it will admit larger problems and be easier to use.

A major extension to the workstation version is also planned. This extension will allow use of auxiliary variables. The algorithms to support this application have been developed and the program structure defined. Generally, the auxiliary variable segment will be a "front—end" to the current program. In other words it will, in effect, be a calling segment for what is now PC CARP.

3. THE U.S. NATIONAL RESOURCES INVENTORY

The Iowa State Statistical Laboratory cooperates with the U.S. Soil Conservation Service on a large survey of land use in the United States. The survey was conducted in 1958, 1967, 1975, 1977, 1982, 1987, and 1992. The survey collects data on soil characteristics, land use and land cover, potential for converting land not used for crops to cropland, soil and water erosion, and conservation practices. The data are collected by employees of the Soil Conservation Service. Iowa State University has responsibility for sample design and for estimation.

The sample is a stratified sample of the nonfederal land area of 50 states and Puerto Rico. The sampling units are areas of land called segments. The segments vary in size from 40 acres to 640 acres. Data are collected for the entire segment on items such as urban land and water area. Detailed data on soil properties and land use are collected at a random sample of points within the segment. Generally, there are three points per segment,but 40—acre segments contain two points and the samples in two states contain one point per segment. Some data, such as total land area and area in roads, are collected on a census basis external to the sample survey.

In 1982, the sample contained about 350,000 segments and nearly one million points. The 1987 sample was composed of about 100,000 segments. The majority of the 1987 sample segments were a subsample of the 1982 segments. Data were collected on about 280,000 points in 1987. The 1992 sample contains about 300,000 segments and about 800,000 points. The majority of the 1992 segments were also observed in 1982.

It is planned to produce a tabulation data set containing data for 1982, 1987 and 1992 for all points in the 1992 sample. This will require imputation for some data for the 1982 and 1987 years for some points. The tabulation data set, together with tabulation software, will be made available to Soil Conservation Service staff at several locations.

The algorithms of PC CARP will form the basis for the estimation software associated with the National Resource Inventory tabulation data set. An interface program is under development. The use will specify the type of table to be constructed, where table is understood to mean estimates for any cross classification. The program will then construct the required variables and call the proper subpopulation option of PC CARP.

ACKNOWLEDGEMENTS

This research was partly supported by Cooperative Agreement 68—3A75—8—12 with the Soil Conservation Service, U.S. Department of Agriculture.

REFERENCES

Binder, D. A. (1983), On the variance of asymp—totically normal estimators from complex surveys. International Statistical Review, 51, 279—292.

Francisco, C. A. and Fuller, W. A. (1991), Quantile estimation with a complex survey design. The Annals of Statistics, **19**, 454—469.

Fuller, W. A. (1975), Regression Analysis for sample survey. Sankhya C, **37**, 117—132.

Fuller, W. A. and Hidiroglou, M. A. (1992), Using Auxiliary Information for a Number of Analyses Options in PC CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.

Fuller, W. A., Kennedy, W. J., Schnell, D., Sullivan, G. and Park, H. J. (1986) PC CARP, Statistical Laboratory, Iowa State University, Ames, Iowa.

Hidiroglou, M. A. (1974), Estimation of regression parameters for finite populations. Unpublishe Ph.D. thesis, Iowa State University, Ames, Iowa.

Hidiroglou, M. A., Fuller, W. A. and Hickman, R. D. (1976), SUPER CARP. Department o Statistics, Iowa State University, Ames, Iowa.

Morel, J. G. (1989), Logistic regression under complex survey designs, Survey Methodology, Vol. 15, 203—223.

Plewis, I. (1989), Review of PC CARP and EV CARP, Applied Statistics, **38**, 529—534.
Sarndal, C. E., Swensson, B. and Wretman, (1992), Model Assisted Survey Sampling, Springer—Verlag.

Schnell, D., Kennedy, W. J., Sullivan, G., Park, H. J., and Fuller, W. A. (1988), Personal Computer Variance Software for Complex Surveys. Survey Methodology, **14**, 59—69.