EVALUATION OF THE EFFICIENCY OF USING PERSONAL COMPUTERS FOR REGRESSION ANALYSIS ON COMPLEX SURVEY DATA

Barbara Lepidus Carlson, Steven B. Cohen, Agency for Health Care Policy and Research
Barbara Lepidus Carlson, 5600 Fishers Lane, Rm. 18A-31, Rockville, MD 20857

KEY WORDS: Statistical software, complex survey design, SUDAAN, PC CARP, regression, logistic regression

The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Health Care Policy and Research is intended or should be inferred. The authors would like to gratefully acknowledge the assistance of several people in carrying out this evaluation: Mr. Gary Moore of Social and Scientific Systems, Inc., for his creation of the data files; Drs. Lee Cornelius and Joel Cohen for providing the analysis models; Mr. Toby Short for hardware and software support; and Ms. Susan Czechowicz for her assistance with the PC runs.

1. INTRODUCTION

Many national surveys have sample designs that deviate from simple random sampling. Stratification is often considered to increase the precision of survey estimates. Clustering is frequently used to make the field work of the survey more efficient. In addition, when greater representation of certain policy-relevant subgroups is necessary, disproportionate sampling is often used. Sampling weights are calculated to reflect the unequal probabilities of selection.

Mainframe computers have been the primary resource used to support federal research and analysis. Since standard statistical computing packages, such as SAS and SPSS, assume simple random sampling, any variance estimates arising from them may not reflect the actual variance achieved by adoption of a more complex design. Specialized software which accounts for complex survey designs when estimating variances has existed for about a decade, but primarily for use on mainframe computers.

With the increased prevalence of personal computers (PCs) and their expanding capacity and speed, the idea of analyzing survey data files on PCs becomes more plausible, particularly moderate-sized data files, say, less than 25,000 records. Using PCs rather than mainframe computers has many potential benefits as well as costs.

The benefits generally relate to cost savings, but freedom from some aspects of mainframe computing also make PCs more attractive. While actual execution time on the mainframe may be substantially less than on a PC, the time from submission of a batch job to the receipt of the printout is generally much longer on a mainframe. Mainframes often operate on a time-sharing basis, which may mean waiting in an execution and/or print queue during a busy period. Due to prohibitive expense, large jobs are often submitted for execution during a discount time, usually overnight or over the weekend, which substantially slows down the entire process. All mainframe computers have down-times, some more often than others. Distribution and installation of tapes on the mainframe is more cumbersome than simply copying diskettes onto a PC. Accessible storage of software as well as data on a mainframe is a daily expense as well.

One generally has immediate access to printouts when using a PC. PC packages often have an interactive or menu format, rather than a batch format, which can make a package easier to learn and use. Furthermore, the freedom from Job Control Language, used to inform the mainframe operating system how to process the job, is an attraction of the PC. There is usually an option with PC packages to output tables and other results from an analysis into a separate textfile, which can be quite helpful later for creating tables without retyping the numbers.

Costs associated with using PCs relate primarily to the run-time issue, as well as space and memory constraints. Obtaining computing equipment well-suited for statistical analysis can become quite expensive, since it must have enough memory, disk space, speed, and often a mathematical co-processor. Unless one has a memory manager which allows for several tasks to be performed simultaneously, a PC can be "tied up" while running a lengthy analysis or downloading a file. The need to download datafiles from the mainframe, along with its potential for introducing transmission errors into the database, is also a consideration.

Weighted regression analysis is widely used with complex survey data for modeling multivariate relationships as well as in the course of nonresponse adjustments and imputation strategies. When computing variance estimates of regression coefficients, specialized software is required in order to make use of the sampling weights and nesting structure. Several such packages exist for mainframe use, but can be quite expensive to run, particularly when several passes are needed to arrive at an optimal model. The cost of weighted logistic regression runs, in particular, can be prohibitive due to the iterative nature of the computations required.

Since large databases require significant resources in order to be analyzed in a timely and efficient manner, under what circumstances will the PC versions of specialized variance estimation programs be useful in keeping computing costs down without sacrificing the efficiency normally associated with the power of a mainframe computer? This is of particular concern with respect to regression analysis. Two of the more frequently-used mainframe packages for the analysis of complex survey data now have PC counterparts: SUDAAN and PC CARP. These PC packages now have the capability of running both weighted least squares and (binomial) logistic regressions. PC CARP can also perform weighted multinomial logistic regressions.

In this paper, the focus is on weighted least squares and logistic regression analysis, and evaluating the feasibility and level of efficiency of using PCs, rather than mainframes, for this purpose. Each of these PC programs is evaluated relative to its mainframe version, and the two PC programs are compared to each other. A comparison solely among the mainframe packages is made as well. Features available in these packages as well as issues related to the actual implementation of the programs, including data preparation steps, number of programming statements, time and cost issues, are examined using a data set from the 1987 National Medical Expenditure Survey (Edwards and Berlin, 1989), which has a complex sample design. Sample programming statements are available from the authors.

2. BACKGROUND

Most of the commonly-used statistical computing packages assume data were obtained from a simple random sample. When data have been collected from a survey which has a complex sampling design, the simple random sample assumption can often lead to an underestimate of the

variance, which can therefore lead to artificially small confidence intervals and anticonservative hypothesis testing; i.e., rejecting the null hypothesis when it is in fact true.

A few different statistical strategies have been developed to address this issue. Among them are: a first-order Taylor series expansion of the variance equation; a balanced-repeated replication method (BRR); and the Jackknife approach (Wolter, 1985). Several software packages have been developed which incorporate one or more of these strategies into their variance calculations.

The current evaluation focused only on those software packages which currently have PC counterparts: SURREGR, RTILOGIT, SUDAAN, and SUPER CARP on the mainframe, SUDAAN and PC CARP on the PC. Other programs which are designed to analyze data from complex surveys exist (OSIRIS PSALMS and OSIRIS REPERR from U. of Michigan, WESVAR and NASSREG from Westat, Inc., and HESBRR from NCHS), but have no PC counterparts to date, and are therefore not pertinent to the subject matter of this paper. These mainframe packages have been evaluated elsewhere (Cohen et al., 1986, Cohen et al., 1988).

SUDAAN (Research Triangle Institute, 1991), SURREGR (Holt, 1977), and RTILOGIT (Shah et al., 1984) are programs appropriate for performing weighted regressions on data from complex survey designs, developed by the Research Triangle Institute (RTI). These programs have many capabilities other than the ones evaluated here. SURREGR, used for weighted least squares regression, and RTILOGIT, used for weighted logistic regression, are mainframe packages only, while SUDAAN has PC, mainframe, and VAX/VMS versions. SUDAAN version 5.50 (April 1991) and PC SUDAAN version 5.41 (February 1991) were used. It should be noted that SUDAAN on the mainframe is still in test mode and is not yet available to the public. Any comments regarding the mainframe SUDAAN should be regarded in this context. SUDAAN will accept both SAS and text datafiles. SURREGR and RTILOGIT will accept only the SAS data format. RTILOGIT can be run only in conjunction with PROC LOGIST, a supplemental SAS procedure (Harrell, 1986).

A Taylor Series approximation is used in SURREGR, RTILOGIT, and SUDAAN to compute variance estimates. RTI has produced a family of such programs, mostly written in SAS language (SAS Institute, Inc., 1985); however, the newest of them, SUDAAN, is written in "C". SUDAAN incorporates the features of SESUDAAN, RATIOEST (ratio estimation package), SURREGR, RTILOGIT, and RTIFREQS (frequencies), and has many improvements over these older versions. Although RTI refers to both the mainframe and PC versions as "SUDAAN", for purposes of clarity, the PC version will henceforth be referred to as "PC SUDAAN" in this paper.

SUPER CARP (Hidirouglou et al., 1980) and PC CARP (Fuller et al., 1988) and its logistic regression supplement (Morel, 1988) are products of the Statistical Laboratory at Iowa State University. SUPER CARP is a mainframe package, the latest version of which is approximately ten years old. PC CARP, its PC counterpart, is relatively recent, and has many improvements over its mainframe parent, and is the only package being evaluated which is run interactively, rather than in batch mode. These programs are written in FORTRAN and also make use of the Taylor Series approximation method. Two supplemental programs, logistic regression and post-stratification, are also available. SUPER CARP and PC CARP will only accept text datafiles. Like the RTI programs, these two packages have many statistical capabilities other than the ones being evaluated here.

3. THE SURVEY DATA

The 1987 National Medical Expenditure Survey (NMES), sponsored by AHCPR, is a national probability sample of the civilian, noninstitutionalized U.S. population. The household survey component was designed to provide statistically unbiased national estimates of health care utilization, expenditures, and access to care, and health insurance coverage for their respective target populations for calendar year 1987. To provide focused estimates of subpopulations of particular policy concern, the Household Survey oversampled the elderly, those with difficulties in performing activities of daily living, poor and low-income families, and the black and Hispanic minorities.

The Household Survey (HHS) sample design can be characterized as a stratified multi-stage area probability design with three stages of sample selection: (1) selection of PSUs (counties or groups of contiguous counties) (2) selection of area segments within PSUs; (3) selection and screening of dwelling units within segments.

To address critical health care policy issues, the economic, sociological, and behavioral studies conducted with NMES data are often characterized by complex multivariate analyses. More specifically, many of these analyses focus on dependent variables that are categorical in nature with two or more classifications.

The application of appropriate logistic or multinomial logistic regression procedures on mainframe computers that adjust for survey design complexities is often characterized by expensive computer runs with charges exceeding $1,000. As a consequence of the frequency of application of these logistic regression analyses for hypothesis testing and estimation of model parameters, and their associated expense, there is great appeal in considering cost-effective analytical alternatives. These concerns served as the motivation for this study, which evaluates the efficiency and analytical capacity of alternative software procedures available for the PC environment.

4. METHODS

The computer packages were evaluated with respect to efficiency, accuracy, and ease of use on both a mainframe and a personal computer, since each package has a version for both environments. For each package, weighted least squares and weighted logistic regression models were estimated for two different models on each of two data subsets: the HHS Medicaid and non-Medicaid populations. The evaluation of the four mainframe and two PC software packages is done by examining and comparing several features for both the weighted least squares and weighted logistic regressions.

The following types of regression models were run. The first type was weighted least squares regression, with a continuous dependent variable: "total number of doctor visits." The second type was weighted logistic regression analysis with a binomial dependent variable: "does the person have a usual source of medical care?" The third and fourth types were weighted logistic regression analyses with multinomial dependent variables of three and four categories: "site of usual source of care" and "type of usual source of care," respectively. All were run with an intercept.

Within each type of regression, two different models were run: a larger model with at least 37 independent variables, and a reduced model, with 12 to 16 independent variables. For each of these models, the analysis was run on two

separate data sets: one containing the Medicaid population and one containing the non-Medicaid population. In addition, the PC runs were done on both a high-speed (33 mHz) and a moderate-speed (20 mHz) computer.

4.1 Computing Environment

The mainframe computer used is an IBM 3090 Model 300J located at the National Institutes of Health in Bethesda, Maryland. It runs under the OS/MVS/ESA operating system. There were two AST brand IBM-compatible personal computers used, both with 80386 processors and 4 mb RAM. One machine used has a 320 mb hard drive (configured as one drive under the MS-DOS version 5.0 operating system) running at 33 mHz and an 80387 33 mHz Intel numeric co-processor. The other machine has a 40 mb hard drive running at 20 mHz (under MS-DOS version 3.3 operating system) and an 80387 20 mHz Intel numeric co-processor.

4.2 Variables

Each of the data sets consisted of stratum and primary sampling unit (PSU) indicators, a sampling weight, and dependent and independent variables on which the regressions were computed. Since the independent variables were primarily (0,1) dummy variables, they were treated as continuous and not specified as categorical variables in the SURREGR, RTILOGIT, SUDAAN, and PC SUDAAN runs.

The files contained observations on respondents to the "Access to Care Supplement" of the Household Survey. Complete documentation on questionnaires and data collection methods is presented in Edwards and Berlin, 1989. Regression models were supplied by NMES analysts. The analytical variables chosen pertained to sociodemographic data as well as issues of access to health care.

The models attempted to predict utilization and usual source of care with independent variables describing age, sex, race/ethnicity, poverty status, health status, education, insurance coverage, family size, region, disability days, and regional health care system descriptors. The multinomial models also had independent variables relating to access to care, functional status, and chronic conditions.

4.3 File Description

The original datafile had 30,038 records. While most independent variables had missing values imputed, records with missing values of model-relevant variables, when they existed, were deleted. This yielded four datasets, two for each subpopulation, with roughly 2,600 records for the Medicaid files and 27,000 records for the non-Medicaid files. In text format, the Medicaid files took up 260 kb. In PC SAS format, these files took up 530 kb. In text format, the non-Medicaid files took up 2770 kb. In PC SAS format, these files took up 5570 kb.

4.4 Procedures

On the mainframe, SAS data sets were created, and SAS (version 5.18) was used to keep only the variables and observations relevant to the evaluation, and delete records with any missing values for those variables. Most missing values were imputed by analysts prior to these runs. However, any remaining missing values were deleted because SUPER CARP and PC CARP do not allow missing values, and the current versions of SUDAAN (5.50) and PC SUDAAN (5.41) have a "bug" regarding missing values, according to the author of the software. For further discussion of the treatment of missing values in SUPER CARP and PC CARP, see Carlson et al. (1990).

To avoid removing any more records than necessary for each of several different models, different subfiles were created based on the variables involved. In addition, the population was divided into Medicaid and non-Medicaid subfiles. For the Medicaid files, eight strata were collapsed into adjoining strata when the subfile yielded only one PSU per stratum. The subfiles were then sorted by stratum and PSU, since all of the programs being evaluated require that the data be sorted by the nesting variables.

Text files were then created from these SAS files using the SAS "FILE" and "PUT" commands. The SAS data files were used for the SURREGR, RTILOGIT, and SUDAAN runs. The text files were used for the SUPER CARP regression runs. Although SUDAAN (and PC SUDAAN) will accept both SAS and text data files, it was decided based on past experience (Carlson et al., 1990) that using the text capability was inefficient in the current version of the package, and was not evaluated here.

On the PC, text files were read into PC SAS (version 6.04, SAS Institute, Inc., 1988), and the same deletions, collapses, and sorts were carried out as on the mainframe. In addition, to save space on the disk, lengths of less than 8 bytes were specified for numeric variables. Subfiles were created and sorted by stratum and PSU. These PC SAS files were used for PC SUDAAN runs. Text files were similarly created from the PC SAS files and were used for the PC CARP runs.

Programming effort was measured by the number of statements required to run the program. When writing the programs to execute the packages being evaluated, an attempt was made to minimize the number of steps needed to execute the program and to make the runs on the various software packages as similar as possible, generally using default options.

Execution times and computing costs, two of the outcomes of interest, were automatically computed and recorded on the printed output from the mainframe runs. Since the analyses were run evenings and weekends, the computing costs were discounted by 60%. These discounted costs are the ones presented here. Had the runs been carried out during prime hours, they would have been two-and-a-half times more costly. A precise execution time for the PC runs is difficult to obtain, as well as inappropriate, due to the different execution modes: the interactive nature of PC CARP versus the batch nature of PC SUDAAN. Therefore, approximate run-times were recorded for the PC runs on both PC CARP (after the last menu prompt) and PC SUDAAN, using the DOS creation time stamp for the output files as the end time. Computational accuracy was evaluated by examining the output from the programs and determining at which decimal place discrepancies began to occur.

5. RESULTS

5.1 Regression capabilities

Weighted least squares regression on the mainframe was carried out using SURREGR, SUDAAN, and SUPER CARP. Weighted logistic regression on the mainframe was carried out using RTILOGIT and SUDAAN. SUPER CARP has no logistic regression capability. Weighted least squares regression on the PC was carried out using PC CARP and PC SUDAAN. Weighted logistic regression on the PC was carried out using PC CARP's supplementary logistic module and PC SUDAAN. Of all the packages being evaluated, only PC CARP has the capability of performing weighted multimonial logistic regression, which was evaluated here, even though no comparison to other packages was possible.

5.2 Programming effort

All of the RTI programs, for both mainframe and PC, were similar in programming effort for both weighted least squares and logistic regressions. (Note that this is true only when

using a SAS data file.) SURREGR and SUDAAN required only five statements, and RTILOGIT required eight statements. As mentioned previously, RTILOGIT must be run in conjunction with the SUGI procedure, PROC LOGIST, whose required statements are counted with the PROC RTILOGIT statements. Programs for PC SUDAAN are identical to the mainframe programs, except, perhaps, for saving output into a file.

While SUPER CARP required only eight statements to run a weighted least squares regression, there is much more effort involved in creating these statements. Numeric codes specifying analyses, variables, and tests are required in specified columns, since the program is written in FORTRAN. The order of the statements in crucial with SUPER CARP, which is generally not the case with the RTI packages.

PC CARP runs in a menu, rather than batch, mode. A series of prompts occur before the analysis is run. The data are read into the program during this phase. For the weighted least squares regression runs, 25 prompts are issued prior to the analysis; for weighted logistic regression runs, 29 prompts are issued. In general, going through the program once gives the user an idea of what types of information concerning the data and the analysis are needed. Subsequent passes can then go through the prompts quickly with the information written out ahead of time.

5.3 Execution time

Comparing execution times between mainframe and PCs is inappropriate, since they run on completely different orders of magnitude. The mainframe execution time is measured in terms of CPU seconds; the PC in terms of minutes. However, as mentioned in the Introduction, the time from the execution of a mainframe job to the receipt of the printout may be far longer than the time elapsed using a PC.

Table 1 shows a comparison of run-times on the mainframe for weighted least squares regression runs. SURREGR ran the fastest of the three packages, with SUPER CARP running two to five times longer, depending on the model and sample size. SUDAAN took much longer than the other packages, and was unable to execute the larger model with the larger sample size in less than 700 CPU seconds. Given the expense, it was decided to quit there.

A similar comparison is shown in Table 1 for weighted logistic regression runs. Of the two mainframe packages with this capability, RTILOGIT ran in significantly less time than SUDAAN, two to almost four times faster. Unfortunately, neither package was able to run the larger model with the larger sample size in less than 900 CPU seconds.

For the mainframe weighted least squares regression runs in all three packages, execution time per observation was the same or slightly less in the larger models; i.e., one can expect a constant increase in execution time per increase in sample size. With respect to size of model, however, execution time per independent variable was one-and-a-half to two times higher in the larger model than in the smaller one using SURREGR and SUDAAN, whereas the time per variable decreased slightly in the larger model using SUPER CARP. Therefore, the cost per variable increases as the number of variables increase in SURREGR and SUDAAN, while there is a slight economy of scale in SUPER CARP.

For the mainframe weighted logistic regression runs in both RTILOGIT and SUDAAN, execution time per observation was about the same in the larger models; i.e., one can expect a constant increase in execution time per increase in sample size. However, execution time per independent variable was one-and-a-half to almost three times higher in the larger model than in the smaller one. Therefore, the cost per variable increases as the model size increases in RTILOGIT and SUDAAN.

Table 2 shows the approximate execution times for the weighted least squares regression runs on the PC. Note that hard disk specifications and software caching can alter these times. The results are meant to show magnitudes of difference. On the higher speed computer, running at 33 mHz, PC CARP executed faster than PC SUDAAN, although not significantly so. In addition, PC SUDAAN was unable to execute either of the larger models with the RAM available, while PC CARP ran the larger models with no problem. Although memory was maximized prior to these runs, neither package makes use of extended memory. Given the few numbers in this table, it appears that there is a slight economy of scale with respect to number of independent variables in PC CARP; i.e., when controlling for the number of variables in the model, the per-variable time is at least 30% less in the larger model. Using the lower-speed computer, running at 20 mHz, the same relationships held, with these runs taking about 60% longer.

One can see the limitations of the PC software in Table 2, the approximate execution times for the weighted logistic regression runs with a dichotomous dependent variable. Only four of the eight attempted models, those with fewer independent variables, ran successfully. PC SUDAAN failed on memory constraints and PC CARP yielded the comment, "The problem specified is too large." For the two that ran successfully on both packages, PC SUDAAN appears to have run in less than half the time on both PCs. Once again, the lower speed computer took about 60% longer.

It was during the PC CARP logistic runs that it was noticed that all of the other packages were able to successfully run weighted least squares and logistic models (for the Medicaid subpopulation) in the presence of two variables with constant values, while PC CARP was not. All of the packages indicated the singularity, but PC CARP quit after a couple of minutes of execution. The PC CARP runs on the Medicaid subpopulation were subsequently re-done with these two problematic variables removed from the model, as were the comparable Medicaid runs on all of the packages. It should be noted that SUPER CARP successfully ran the weighted least squares regressions with this singularity.

The runs that ran out of memory in PC SUDAAN took some time as well. For the larger weighted least squares regression models, the programs for the Medicaid subpopulation and the Non-Medicaid subpopulation ran for 2 and 24 minutes, respectively, before issuing the error message. For the larger weighted logistic regression models, the programs for the Medicaid and Non-Medicaid subpopulations ran for 14 minutes and 135 minutes, respectively, converged after six iterations, and then issued the error message. For these same logistic models, PC CARP took no time before telling the user that the problem was too big, presumably due to memory constraints.

The time PC CARP uses to read in the data during the interactive prompts is negligible, perhaps a few seconds. PC SUDAAN reads in the observations multiple times for each analysis, with a counter appearing on the screen. For the weighted least squares runs, it counts through the observations twice. For the weighted logistic runs, it counts through the observations initially and then recounts for each iteration until it converges. It is at this point that PC SUDAAN crashed due to lack of memory. This seems an inefficient way to input and store data, in contrast to PC CARP.

Weighted logistic models with dependent variables of three and four categories were also tested in a similar manner. Only the two models with the smaller sample size and a three-category dependent variable ran successfully, and only one of those ran on the slower PC. Neither package was able to handle weighted multinomial logistic regressions with a dependent variable of four categories, given the sample sizes and model sizes evaluated here.

5.4 Computing costs

There are no costs associated with running software on a PC, other than initial purchase costs of hardware and software. On the mainframe, costs are determined by a combination of factors, including CPU time. Therefore, the relationships among the packages reflected in the previous section, Execution time, exist with respect to cost as well. One must keep in mind that the costs reported are discounted by 60% because the jobs were issued during evening and weekend hours. If they had been run during prime hours, the costs would have been two-and-a-half times greater. The discounted costs for these runs were not insignificant for the weighted least squares regression runs on the larger data set, ranging from $5 to $135, with the SUDAAN run unable to run in the allotted time, yet costing over $500.00.

The situation was worse with the weighted logistic regression runs on the mainframe. Only the smaller model run on the smaller data set was less than $15. The ones that ran ranged in cost from $4 to $215. The jobs which were unable to run also cost more than $500.00 each.

5.5 Computational accuracy

The estimates of regression coefficients and their standard errors from the various packages were compared. The default number of significant digits in the outputs differed from program to program; therefore, the comparisons were made out to the minimum number of decimal places in common. In SURREGR and RTILOGIT, general mean square errors are presented, rather than standard errors; therefore, their square root was taken for comparison purposes.

In comparing SURREGR to SUPER CARP, the estimated regression coefficients were exactly the same, and their standard errors converged out to three places beyond the decimal point. Comparing SUDAAN (and PC SUDAAN) to SUPER CARP (and PC CARP) yielded the same coefficients out to at least two decimal places, and the same standard errors out to at least one decimal place. This level of convergence is not surprising given that all of the packages evaluated use the Taylor approximation to compute variances. It should be noted that the PC handled the same desired level of precision as was acquired on the mainframe.

One curious result occurred with the weighted logistic regression run for the smaller model on the non-Medicaid subpopulation. While SUDAAN and PC SUDAAN do not indicate the number of iterations until convergence on their output files, the RTILOGIT converged in six iterations, while PC CARP converged in five iterations. In addition, PC CARP's regression coefficients, in this case only, had opposite signs than those for SUDAAN, PC SUDAAN, and RTILOGIT. These two disparities did not occur for the other models.

5.6 Quality of documentation

The relative ease with which one learns and uses a statistical computing package is a function of prior computing experience, statistical background, and the clarity and comprehensiveness of the documentation. Manuals (with or without on-line help facilities) should provide enough information so that the first-time user can learn the package without assistance. Error messages should be explicit enough so that the user can understand and correct the problem. Unlike the other measures used to compare the software packages, evaluation of documentation is somewhat subjective and can vary between users.

For the most part, the software documentation was quite good for all of the packages being evaluated. Examples are used to some extent in all of the manuals (except for SURREGR), and are quite helpful when one is using one of the packages for the first time. All contain algorithms for the available analyses, for those interested in the technical aspects. The documentations evaluated for both SURREGR and RTILOGIT were perhaps not the final versions.

The SURREGR manual is concise, but adequate. It has relatively clear instructions on how to structure the program statements; however, there are no examples given, which would be useful. The RTILOGIT manual has a technical section as well as a large sample program/output section. The instructions on how to structure the program statements are found in its appendix.

The SUDAAN manual is designed for use with the PC and the VAX, not for the mainframe. Although the program commands are the same for the mainframe, a mainframe section would need to be added to the manual to give some information on Job Control Language and file-naming, and other details related to the interaction with the mainframe system. The existing SUDAAN manual, for its intended environment, is organized and clearly-written. It should be noted that the R^2 and number of iterations are shown only on the screen for PC SUDAAN and not in the output file.

The SUPER CARP manual, although clearly written, is quite dated. It is written in terms of "punched cards," which can be interpreted as lines of code. However, it was not initially clear that one cannot have a blank line, for example. The manual presumes some prior knowledge of FORTRAN and its data formats, as well as the way it reads data files. Some attention should be paid to updating the SUPER CARP manual, since it still is useful in cases where a file is too big for the PC to handle efficiently and effectively. Some information on SUPER CARP's interaction with the mainframe (e.g., file-naming/numbering) should also be added.

The PC CARP manual is very well-written and demonstrates how to use the package primarily through the use of examples. Screen displays are shown throughout the examples. There is also an on-line help feature in PC CARP. The documentation for the logistic regression supplement is a bit heavy on the technical aspects and light on examples and other information which might be helpful. For example, no mention is made of restrictions on the number of variables to be input into a model, or the fact that the dependent variable can have more than two categories (one of the package's strengths).

5.7 Miscellaneous flaws and limitations

With respect to weighted regression analysis, SUDAAN has many advantages over its predecessors, SURREGR and RTILOGIT, although some flexibility was lost in the transition. The older packages had the benefit of being a procedure within a SAS program. Data could be easily manipulated within the same program using SAS DATA steps, whereas the dataset running under SUDAAN has to be a permanent data set, making it more difficult to modify variables and re-run the program.

However, programming a weighted logistic regression run in SUDAAN is more straightforward than using RTILOGIT,

since there is no need to first run the supplementary SAS procedure, PROC LOGIST. SUDAAN also allows for post-stratified estimates and has much more flexibility with respect to the sampling design. (With-replacement was assumed here.) In addition to the without-replacement design, the first stage can be specified as without-replacement for either a simple random sample or unequal probability of selection. Later stages can be specified as with or without replacement. A very helpful feature of SUDAAN is that the statements are identical on the mainframe and the PC. Once the JCL is "mastered" on the mainframe, the rest of the SUDAAN specifications are straightforward.

PC CARP has several improvements over its predecessor, SUPER CARP, most prominently its ability to do weighted logistic regression analysis, as well as other new features. In addition, SUPER CARP requires that data be specified in particular columns, in a somewhat scattered set pattern. The order of the rows is not at all flexible. In PC CARP, that is not an issue, due to the menu-driven mode of specification. Both packages have a maximum limit on the number of variables to be input, not necessarily analyzed, set at 50. Neither package provides p-values for its test statistics, unlike the other programs being evaluated.

PC CARP and PC SUDAAN are comparable in their space and memory requirements. The PC SUDAAN software takes up roughly 1 mb of disk space, and requires 640 kb RAM. It can run on any IBM-compatible PC. The PC CARP software takes up roughly 470 kb (including the logistic regression supplement), and requires 450 kb RAM. It can run on an IBM-compatible machine with a mathematical co-processor. Neither package makes use of extended memory. The two packages are roughly equivalent in cost ($350-$500).

Limitations of PC CARP are that it can only read in textfiles, and its inability to run with missing data. Many of the difficulties found with both SUPER CARP and PC CARP are due to the fact that the programs are written in FORTRAN: the specification of input format, the output expressed in scientific notation, the rigid column format with SUPER CARP. The interactive format can be seen as a feature or a fault, depending on one's preference. It is sometimes difficult to change a response or get out of an erroneous keypunch within PC CARP. A batch option would satisfy those who prefer that method of execution.

A note here about using the PC for data manipulation. PC SAS was used to create dummy variables, sort the files, create textfiles, and otherwise handle the data prior to execution of these programs. A large amount of disk space is needed to store PC SAS, and even more space is needed for workspace when executing SAS. Under DOS version 3, the size of configured drives is limited. Due to the size of the data files used for this evaluation, a good deal of "file shuffling" was needed to do any of the PC SAS jobs, even on the PC with 320 mb of hard disk space. With DOS version 5, this restriction no longer applies, and so the hard disk was configured as one drive, allowing full use of the hard disk space for PC SAS execution.

6. SUMMARY

Using data from the National Medical Expenditure Survey, six widely used regression programs (multivariate or logistic regression) appropriate for the analysis of complex survey data were compared. The four programs developed for mainframe computers under investigation included: SURREGR, SUDAAN, RTILOGIT, and SUPERCARP. In addition, the two programs developed for analysis on the PC under investigation were PC SUDAAN and PC CARP. Particular attention was directed to a comparison of the efficiency of using statistical software developed for personal computers as an alternative to their mainframe counterparts when conducting the same multivariate analyses. The comparisons also concentrated on user facility, computational efficiency, computational accuracy, quality of documentation, and program limitations. The study was also designed to measure the effect of alternative specifications for database size and number of independent predictor variables on program performance.

As a consequence of the frequency of application of multivariate regression and logistic regression analysis for NMES analytical reports, the identification and subsequent use of the most efficient software procedure within a personal computing environment should yield substantial savings in survey costs. In this analysis, it was determined that both the PC SUDAAN and PC CARP software packages are viable alternatives to their mainframe counterparts for NMES analyses when the number of predictor variables under consideration is constrained to thirteen or less. When much larger multivariate prediction models are under consideration (e.g. over 30 independent predictors), memory constraints within the PC environment seriously limit the performance of these software procedures.

The key problem with the PC packages evaluated in this paper is their inability to make use of extended memory, or to even page some memory off onto the hard disk. Analysts are more than willing to let a program run slowly overnight on a PC, as long as they know their output will eventually be there, in order to avoid the high mainframe computing costs. These two PC packages are running into memory problems on computers with vast memory (4 mb) and hard disk (320 mb) capacities. In fact, these programs crashed under the most favorable conditions with respect to executable memory: each PC had more than 500 kb RAM available for use, and this was accomplished by the use of memory management software which enabled us to free up execution space. The average user may run into problems with much smaller files and models.

(For copies of the reference section, please contact the authors at (301) 443-4836, or in writing.)

Table 1 Execution Times (CPU seconds) for Mainframe Weighted Regression

	Least Squares			*Logistic*	
	SUR-REGR	SUDAAN	SUPER CARP	RTI-LOGIT	SUDAAN
Smaller Model	(13 indep. variables)			(12 indep. variables)	
Medicaid (n=2,585)	0.89	16.52	3.80	7.52	28.02
Non-Medicaid (n=27,386)	8.48	191.91	41.56	120.57	307.49
Larger Model	(38 indep. variables)			(37 indep. variables)	
Medicaid (n=2,585)	5.19	95.73	9.55	73.23	136.10
Non-Medicaid (n=27,386)	34.76	*	102.13	**	**

* abended at 700 CPU seconds ** abended at 900 CPU seconds

Table 2 Approximate Execution Times (in minutes) for PC Weighted Regression Runs (on 386/33 PC)

	Least Squares		*Logistic*	
	PC SUDAAN	PC CARP	PC SUDAAN	PC CARP
Smaller Model	(13 indep. variables)		(12 indep. variables)	
Medicaid (n=2,585)	2	1	3	7
Non-Medicaid (n=27,386)	19	13	39	83
Larger Model	(38 indep. variables)		(37 indep. variables)	
Medicaid (n=2,585)	*	2	*	**
Non-Medicaid (n=27,386)	*	21	*	**

*"Not enough memory for this job" **"The problem specified is too large"
N.B. Hard disk specifications and software caching can alter these times.