Journal of Official Statistics
Vol. 4, No. 1, 1988, pp. 17–34
© Statistics Sweden

An Evaluation of Statistical Software Procedures Appropriate for the Regression Analysis of Complex Survey Data

Steven B. Cohen,[1] *Judy A. Xanthopoulos,*[1] *and Gretchen K. Jones*[2]

Abstract: Data from complex survey designs require special consideration with regard to variance estimation and analysis, because of design components that include unequal selection probabilities, stratification, and clustering. Statistical software package programs are currently available which accommodate a complex survey design, and allow for the generation of centrality parameters and variance estimates for statistics expressed in terms of means, totals, ratios, and multivariate regression coefficients. The methods of variance estimation include the Taylor series linearization method and balanced repeated replication. Using data from the National Medical Care Expenditure Survey, which is characterized by a highly complex survey design, the following four statistical programs appropriate for multivariate analysis of complex survey data are compared: SURREGR, SUPERCARP, REPERR, and NASSREG. The comparison focuses on cost-efficiency, user facility, and program capabilities for a series of regression analyses that are representative of the analytical requirements of the National Medical Care Expenditure Survey.

Key words: SURREGR; SUPERCARP; REPERR; NASSREG.

1. Introduction

Data from complex survey designs require special consideration with regard to variance estimation and analysis, as a consequence of a departure from the assumption of simple random sampling. National surveys conducted by government organizations, industry, political organizations, and market research firms, often with disparate intentions, share a determination to provide the greatest precision in estimates from sample data for fixed cost and time constraints. Consequently, many national surveys are characterized by design components which include stratification, clustering,

[1] National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD 20857, U.S.A.
[2] National Center for Health Statistics, Hyattsville, MD 20782, U.S.A.

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 National Center for Health Services Research and Health Care Technology Assessment is intended or should be inferred.

The authors wish to thank Dr. David Fitch for his statistical support, Mr. Samuel Meyers, Dr. Alan Monheit, and Ms. D.E.B. Potter for their helpful comments, and Mr. William Ashby for his skillful typing of this manuscript.

and disproportionate sampling. Statistical software packages are currently available which accomodate a complex survey design. They can generate variance estimates of statistics expressed in terms of means, totals, ratios, and regression coefficients. The procedures vary, however, in terms of program capabilities, computational efficiency, and user facility.

Using data from the National Medical Expenditure Survey, which is characterized by a highly complex survey design, existing statistical programs appropriate for the regression analysis of complex survey data are compared. The programs under investigation are: the SURREGR procedure developed by the Research Triangle Institute (Holt (1982)), the SUPERCARP program developed by the Statistical Laboratory at Iowa State University (Hidiroglou, Fuller, and Hickman (1980)), the REPERR procedure in OSIRIS IV, developed at the University of Michigan (Van Eck (1979)), and the NASSREG procedure developed by Westat (Chu, Mohadjer, Morganstein, and Rhoades (1985)). The comparison focuses on analytical flexibility, computational efficiency, and user facility for a series of multivariate analyses that are representative of the analytical requirements of the National Medical Care Expenditure Survey.

2. Background

Many general purpose national sample surveys adopt stratification as a design feature to increase the precision of survey estimates. Disproportionate sampling is another strategy adopted to insure sufficient representation of specific subgroups from an underlying population, while simultaneously allowing for the capacity to yield precise estimates of relevant characteristics for the complete target population. Cluster sampling is a third method used in combination with stratification and probability selection schemes to increase a sample's efficiency. Other departures from equal selection probability sampling in large scale surveys are due to coverage deficiencies. Generally, specific subgroups of an underlying population are more prone to refuse participation in the survey, forcing consideration of nonresponse adjustments to minimize bias when estimating relevant population parameters and totals. All these methods of sampling, as they depart from an equal probability selection design, are components of the set of survey strategies which create a complex survey design. Further complexities are added to the estimation process through the derivation of sampling weights, which reflect unequal selection probabilities and include poststratification adjustments (Cohen (1983)).

Due to the nature of complex survey designs, there is cause for concern regarding the methods of variance estimation and the subsequent use of the estimates in the construction of confidence intervals and hypothesis testing. Standard methods of variance estimation for means, proportions, ratio estimates, and regression coefficients, which are present in the most commonly used statistical package programs, assume simple random sampling (such as SPSS (Statistical Package for the Social Sciences), SAS (Statistical Analysis System), BMDP (Biomedical Data Program) and OSIRIS). When this approach is directly used with data from a complex survey design, the result is often an underestimate of the true variability of a statistic (Cohen and Kalsbeek (1981)). The consequences for statistical inference is an anticonservative test. This suggests that the probability of rejecting the null hypothesis is greater than expected when an appropriate estimate of variance is used.

Several methods for approximating sampling variances, which incorporate the components of a complex survey design, have been developed. The three most generally accepted and frequently used techniques are the method of Balanced Repeated Replication

(BRR), the "jackknife" method, and the Taylor series linearization method (Cohen and Kalsbeek (1981)). These variance estimation strategies have been incorporated as procedures in several of the widely used statistical packages.

2.1. *Population estimates*

When analytical attention is directed towards statistics expressed in terms of means, proportions, totals, and ratios, unbiased population estimates and their respective standard errors can be derived by use of the following programs: SUPERCARP, SESUDAAN/RATIOEST, PSALMS, or HESBRR. The SUPERCARP program (Cluster Analysis and Regression Program), developed by the Survey Section of the Statistical Laboratory at Iowa State University, uses the Taylor series linearization method of variance estimation appropriate for complex survey data. The SESUDAAN program (Standard Errors Program for Computing of Standardized Rates from Sample Survey Data), which is accessible through SAS, also allows for the generation of variance estimates for means, totals, and proportions using the Taylor series linearization method (Shah (1981a)). Similarly, variance estimates of ratios can be generated through the SAS accessible procedure, RATIOEST (Standard Errors Program for Computing of Ratio Estimates from Sample Survey Data), also using a linearization approximation (Shah (1981b)). The PSALMS (Sampling Error Analysis) procedure is a component of the OSIRIS IV Statistical Analysis and Data Management Software Systems Package (Van Eck (1979)), and also considers the Taylor series linearization method of variance estimation for complex survey data. Alternatively, the balanced half-sample method of variance estimation can be implemented through the Health Examination Survey Variance and Cross Tabulation Program (HESBRR) developed by the National Center for Health Statistics (Jones (1983)). A comparison of the performance of these four variance estimation programs was conducted for survey statistics expressed in mean, total, and ratio form using data from a complex National Medical Care Expenditure Survey. Study findings revealed the SESUDAAN/RATIOEST procedure to be the most efficient program in terms of Central Processing Unit (CPU) time used, and consistently superior in terms of programming facility (Cohen, Burt, and Jones (1986)).

2.2. *Multivariate analysis*

Research hypotheses that focus on the determination of the relationship between relevant health care measures and potential predispositional factors, fall within the framework of multivariate regression analysis. The general multiple regression model takes the form:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \ ,$$

where

$\mathbf{Y}$ is an n by 1 vector representing sample observations of the dependent variable,

$\mathbf{X}$ is an n by p matrix of sample data for the (p-1) predispositional variables and intercept term,

$\mathbf{B}$ is a p by 1 vector of parameters to be estimated, and

$\mathbf{E}$ is an n by 1 vector of error terms.

The classical assumptions associated with the model are:

- The **X** matrix is composed of nonstochastic terms. In addition, no exact linear relationship exists among two or more of the independent variables.

- The dependent variable is normally distributed with mean **XB** and constant variance $\mathbf{I}\sigma^2$, where **I** is an (n by n) identity matrix.

- The vector **E** consists of independent, identically distributed error terms which follow a normal distribution with zero mean and constant variance $\mathbf{I}\sigma^2$.

When a sample design employs a scheme with varying selection probabilities, parameter estimates for the regression model must be derived using weighted least squares. Consequently, the matrix of estimated model parameters, $\hat{\mathbf{B}}$, takes the form:

$$\hat{\mathbf{B}} = (\mathbf{X}'\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{W}\,\mathbf{Y},$$

where **W** is an n by n diagonal matrix of analysis weights associated with the n sample observations, reflecting their probabilities of sample selection.

The properties of the parameter estimates derived from weighted least squares estimation are discussed in Holt, Smith, and Winter (1980), Kish and Frankel (1974), and Shah, Holt, and Folsom (1977). Survey design complexities necessitate that variances of the estimated model parameters be estimated by either jackknife, balanced replication, or Taylor series linearization methods. Hypothesis testing for model parameters can then be conducted by application of t-tests for individual parameters, of F-tests for multivariate considerations.

Several regression programs are currently available that are appropriate for the analysis of complex survey data. The SUPERCARP procedure incorporates the weighted least squares method of regression coefficient estimation, in addition to computing variances of the estimated coefficients by a Taylor series approximation, to accommodate survey design complexities. These features are shared by a related regression program, SURREGR (Standard Errors of Regression Coefficients from Sample Survey Data (Holt (1982)), which is accessible through SAS. Alternatively, one can gain access to repeated replication procedures to compute standard errors of regression coefficients derived from complex survey data through the OSIRIS IV Repeated Replication Sampling Error Analysis procedure, REPERR. This procedure allows for the creation of replications using one of three methods: balanced half-sample, jackknife, or user specified replications. In addition, the SAS accessible NASSREG procedure developed by Westat (Chu et al. (1985)), also uses the balanced half-sample replication technique for variance estimates of regression coefficients.

3. Analytical Requirements of the National Medical Care Expenditure Survey

The National Medical Care Expenditure Survey (NMCES) was conducted to meet the needs of government agencies and health professionals for more comprehensive data on the utilization, costs, and sources of payment associated with medical care in the United States. The survey, which was cosponsored by the National Center for Health Services Research (NCHSR) and the National Center for Health Statistics (NCHS), has a complex design. A stratified multistage area probability design was further complicated by combining two independently drawn national samples of households, one by the Research Triangle Institute (RTI) and one by the National Opinion Research Center (NORC). Sampling specifications called for the selection of approximately 14 000 households to represent the civilian noninstitutionalized population, with six interviews over an 18 month period during 1977 and 1978. The survey was complemented by additional surveys of physicians and health care facilities providing care to household members during 1977 and of employers and insurance companies responsible for their insurance coverage.

Topics of particular interest to government agencies, legislative bodies and health professionals include:

> The cost, utilization, and budgetary implications of changes in federal financing programs for health care and of alternatives to the present structure of private health insurance.
>
> The breadth and depth of health insurance coverage.
>
> The proportion of health care costs paid by various insurance mechanisms.
>
> The influence of Medicare and Medicaid programs on the use and costs of medical care.
>
> How and why Medicaid participation changes over time.
>
> Patterns of use and expenditures as well as sources of payment for major components of care.
>
> The cost and effectiveness of different federal, state, and local programs aimed at improving access to care.
>
> The loss of revenue resulting from current tax treatment of medical and health insurance expenses, particularly with regard to the benefits currently accruing to different categories of individuals and employers, and the potential effects on the federal budget of proposed changes to tax laws.
>
> How costs of care vary according to diagnostic categories and treatment settings.

To address these critical health care policy issues, the economic, sociological, and behavioral studies conducted with NMCES data were often characterized by complex multivariate analyses. To date, over 100 NCHSR analytical papers have been written using the NMCES data. As a consequence of the frequency of application of multivariate regression analysis for hypothesis testing and estimation of model parameters, a study of the efficiency and analytical capacity of alternative software procedures appropriate for complex survey data was considered. It was believed that the identification and subsequent use of the most efficient software procedure, in terms of computer time and user facility, would yield substantial savings in survey costs associated with data analysis.

3.1. Study design

For the purposes of this study, a representative set of health care utilization, expenditure, and morbidity measures were specified as criterion variables to typify the economic, sociological, and behavioral multivariate analyses conducted with the NMCES data. The utilization measures included the number of outpatient physician contacts, hospital admissions, dental visits, and the number of prescribed medicines. More specifically, all outpatient physician contacts made in 1977 included telephone calls. Hospital admissions included admissions of less than 24 hours and those for women giving birth. Newborns were not counted as separate admissions unless they were admitted separately following delivery. Dental visits included all visits to a dentist, dental surgeon, oral surgeon, orthodontist, other dental specialist, dental hygienist, dental technician, or any other person for dental care. Prescribed medicines included any drug or other medical preparation prescribed by a physician, including refills. Expenditure data for each of these specified utilization measures was also considered. Disability days served as the measure of morbidity, which included the number of days illness or injury kept a person in bed, away from job or other work, or usual activity (e.g., work around the house, school).

Since the selected variables were primarily measures of health care demand and morbidity, the model specification in this study incorporated a set of explanatory variables consistent with the demand equation specifications in the health economics literature (Newhouse and Phelps (1976)). The following explanatory variables were included in the respective regression models: age, sex, race (white, nonwhite), health status (excellent/good, fair/poor), size of city (SMSA, non-SMSA), region, education of household head (college graduate, other), health insurance coverage (ever insured, never insured), private insurance coverage (ever covered, never covered), Medicaid coverage (ever covered, never covered), and family income (Table 1).

Table 1. Regression model specification

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \ldots + B_{13}X_{13i} + e_i$$

$$i = 1, 2, \ldots, n$$

Dependent variables

Y–1. Number of outpatient physician contacts
Y–2. Expenditures for outpatient physician contacts
Y–3. Number of hospital admissions
Y–4. Expenditures for hospital admissions
Y–5. Number of dental visits
Y–6. Dental visit expenditures
Y–7. Number of prescribed medicines
Y–8. Prescribed medicine expenditures
Y–9. Number of disability days

Independent variables

X_1 = SEX (1 = Male, 0= Female)
X_2 = AGE
X_3 = HEALTH STATUS (1 = Excellent/Good, 0 = Fair/Poor)
X_4 = SMSA STATUS (1 = SMSA, 0 = NonSMSA)
X_5 = RACE (1 = White, 0 = Other)
X_6 = EDUCATION OF HOUSEHOLD HEAD (1 = College grad, 0 = Other)
X_7 = HEALTH INSURANCE COVERAGE (1 = Never covered, 0 = Ever Covered)
X_8 = PRIVATE INSURANCE COVERAGE (1 = Ever covered, 0 = Never Covered)
X_9 = MEDICAID COVERAGE (1 = Ever covered, 0 = Never covered)
X_{10} = FAMILY INCOME
X_{11} – X_{13} = REGION INDICATORS
X_{11} – (1 = Northeast, 0 = Other)
X_{12} – (1 = North Central, 0 = Other)
X_{13} – (1 = South, 0 = Other)

Regression analyses

Single analysis – Y–2
Five analyses – Y–1, Y–2, Y–3, Y–8, Y–9
Nine analyses – *All*

For a given model specification, the software procedures under investigation were required to: (1) produce regression coefficient estimates, (2) generate the related variances or standard errors of the coefficient estimates, and (3) perform tests of significance for the individual model coefficients and overall model specification. The software comparison was primarily directed to program performance in terms of computational efficiency and user facility. The study was broadened to consider the effect of sample size on computational efficiency.

The complete NMCES data base contains observations on 38 815 individuals. Often, descriptive and analytical reports concentrate on only a subset of that data base (e.g., the poor, the elderly, the uninsured, the unemployed). In addition, other data sets employ a complex survey design similar to the NMCES, but are based on smaller samples. For example the National Medical Care Utilization and Expenditure Survey (NMCUES), conducted by NCHS in 1980, used a sample design and questionnaire similar to the NMCES. However, the sample included a much smaller number of participants (17 123 individuals). Consequently, the software comparison concentrated on three distinct NMCES samples of different sizes. The intent was to generalize study findings beyond the scope of the NMCES. The three samples of interest were: the overall NMCES sample of 38 815 individuals, the 8 350 individuals in NMCES greater than or equal to 55 years of age, and the 1 120 individuals in NMCES that were unemployed.

Further, the specified number of regression analyses in a given production run were varied, to determine whether economies of scale were achievable. Three distinct production runs were considered: a single regression analysis, five regression analyses, and nine regression analyses. The nine regression analyses were distinguished by the different utilization, expenditure and morbidity measures specified as criterion variables for the study. The remaining production runs considered subsets of the nine distinct model specifications. The final design of the experiment could be characterized by a 3x3 table, whose marginals were represented by the three distinct classes of the two factors: sample size and number of regression analyses (Table 2).

Table 2. Study design

Number of regression analyses	Sample size $n = 38\ 815$	 $n = 8\ 350$	 $n = 1\ 120$
1	SR, SC, RP, NR	SR, SC, RP, NR	SR, SC, RP, NR
5	SR, SC, RP, NR	SR, SC, RP, NR	SR, SC, RP, NR
9	SR, SC, RP, NR	SR, SC, RP, NR	SR, SC, RP, NR

SR denotes the SURREGR procedure.
SC denotes the SUPERCARP procedure.
RP denotes the REPERR procedure.
NR denotes the NASSREG procedure.

Source: National Center for Health Services Research and Health Care Technology Assessment.

4. Software Comparisons

The four software procedures, SURREGR, SUPERCARP, REPERR, and NASSREG, were compared with respect to ease of application, computer processing time, and program capabilities. Prior to execution of the procedures, the input data sets were sorted by two variables which identified sampling levels of the NMCES. Since the NMCES had a multi-stage sample design, the identification of the sampling unit levels was required for estimating standard errors. The first stage sampling level variable identified the respective strata of the design. The primary sampling units (e.g., counties) selected within strata were designated by the second stage sampling level variable.

4.1. *User facility*

The SURREGR procedure required only five program statements (Table 3) to produce the estimated regression coefficients, related standard errors, and required tests of significance for model coefficients, for the complete set of regression analyses presented in Table 1. The PROC statement identifies the SURREGR procedure, and invokes the computation and printing of estimated regression coefficients. The PSU and STRATUM statements identify the nested structure of the sample design and the WEIGHT statement specifies the sampling weight. The MODEL statement identifies the dependent variable(s) and the specified explanatory variables. Within this framework, the procedure allows for multiple regression analyses with the same set of independent factors by specifying a set of distinct dependent variables in the model statement. Consequently, five program statements were also needed for the production run with five regression analyses specified, and for the single regression analysis. Alternatively, only the model statement needs to be invoked for each additional model specification that alters the set of explanatory measures.

Table 3. Ease of application – Number of programming statements

Number of	Statistical software procedure			
regression analyses	SURREGR	SUPERCARP	REPERR	NASSREG
1	5	8	5	3
5	5	24	5	15
9	5	40	5	27

Source: National Center for Health Services Research and Health Care Technology Assessment.

The SUPERCARP procedure required forty program statements (Table 3) to yield the required output for the complete set of nine regression analyses. The first program statement is the PARAMETER statement which requires specification of a program title, the sample size, stratum sampling rates, the number of input variables, and the input file number. The parameter statement also indicates whether a counter variable is requested, whether a complex survey design is considered, the total number of regression analyses, listing options for sample data, and whether strata which contain only one observation are to be collapsed. The second program statement is referred to as the FORMAT statement, and specifies the input format of the data set. The VARIABLE NAME statement is the third required program statement, and it contains names for all input variables. The next program statement is referred to as the SCREENING OPERATION statement, and allows for the removal of sample observations in subsequent analyses. The ANALYSIS statement then identifies the type of analysis to be considered, the number of variables considered, the type of regression analysis, an intercept indicator, and an indicator for the number of tests on groups of regression coefficients that will be performed. A VARIABLE IDENTIFICATION statement then identifies the dependent and independent variables to be considered in the regression analysis. Two additional program statements are needed to test the overall model specification, indicating the number of coefficients to be tested and their identification. The four program statements which constitute a regression analysis specification must be repeated for each distinct regression analysis under consideration. Consequently, eight program statements are required for a single regression analysis specification and twenty-four program statements are necessary for the five regression analyses under consideration.

The OSIRIS IV REPERR procedure required five program statements (Table 3) to produce the estimated regression coefficients, related standard errors, and other required output for the complete set of regression analyses. The &REPERR statement calls the REPERR procedure in OSIRIS and identifies the input data set and its related dictionary. The next statement allows for a user specified label for program output. A parameter program statement follows, which identifies the sampling strata variable (STRATA=) and the sampling error computation unit variable (SECU=), the sampling weight (WTVAR=), the statistics to be calculated (STAT=), and the type of printed output (REGR=). A model statement is then required, which identifies the list of stratum values to be used in the analysis (STRATA=), and the replication formation model to use (MOD=). In this study, the balanced half-sample method of variance estimation was considered. Finally, a regression statement is required to identify the independent variables (VARS=) and dependent variables (DEPV=) for the specified regression analyses.

Similar to the SURREGR procedure, REPERR allows for multiple regression analyses with the same set of independent factors, with the specification of a set of distinct dependent variables in the regression statement. Five program statements were also required for the production runs with five regression analyses specified, and for the single regression analysis.

The NASSREG procedure required twenty-seven program statements to yield the desired program output for the complete set of nine regression analyses. A minimum of three program statements is required per regression analysis. The PROC statement identifies the NASSREG procedure, and invokes the computation and printing of the estimated regression coefficients and standard errors. The MODEL statement identifies the dependent variable and the associated independent variables for a specific regression analysis. The WEIGHT statement identifies the full sample weight followed by the set of half-sample weights required to compute variance estimates of regression coefficients by the method of balanced repeated replication. Only one dependent variable may be used in a MODEL statement, and only one MODEL statement may be used for each PROC statement. These three statements must be repeated for each separate regression analysis under consideration.

As a SAS accessible procedure, the SURREGR program is user friendly and requires the minimum number of programming statements for the complete set of analyses under consideration. NASSREG is also a SAS accessible procedure, requiring only three program statements for a single regression analysis. However, the user must supply the required set of replicate weights as program input for variance estimation using the method of balanced repeated replication. The OSIRIS IV REPERR procedure is also straightforward in its application, and requires the equivalent number of programming statements characterizing SURREGR, for the specified regression analyses. Unlike NASSREG, this program directly generates the required set of half-sample weights for variance estimation. The SUPERCARP procedure requires more programming interaction to implement, both in terms of program statements and attention to detail. The program format is quite disciplined, where program commands must be specified in fixed column locations. The SURREGR and NASSREG procedures operate on a SAS data set, whereas the SUPERCARP procedure inputs data stored in fixed block format through Fortran specifications. As noted, the REPERR procedure requires an OSIRIS IV data set for program input.

4.2. *Program efficiency*

The alternative software procedures were then compared in terms of computational efficiency. The SURREGR, SUPERCARP, and NASSREG procedures were run on the National Institutes of Health (NIH) Computer System, which is an IBM 370 facility. The NIH Computer System is located on the NIH Campus at 9000 Rockville Pike, Bethesda, Maryland 20205. Due to accessibility considerations, the OSIRIS IV REPERR procedure was run on the Parklawn Computer System, which is an IBM 370/3081 facility. The Parklawn Computer Center is located at 5600 Fishers Lane, Rockville, Maryland 20857.

The comparisons focused on the Central Processing Unit (CPU) time that was required to run the respective regression analysis programs. A direct measure of computer charges was not considered in the comparison due to dramatic variations in charging algorithms across installations. These charging algorithms are generally a function of CPU time, region size, number of tape drives, and Input/Output processes. Since computation cost is directly related to CPU time used for a particular job, this measure served as an indicator of both computational efficiency and computer cost.

As noted, the alternative regression analysis programs were not all currently supported by a single computer installation that was accessible. Consequently, a standardized measure of CPU time had to be developed to facilitate comparisons across installations. This was achieved by running the SURREGR procedure at both facilities. The data set representing the NMCES sample of individuals aged 55 and over (n=8 350) was used, and the three distinct regression runs which varied the number of regression analyses specified in Table 1 were compared in terms of CPU time. The ratios of the NIH computer CPU time to the Parklawn computer CPU time for the three distinct standard error runs were: 0.42 for a single regression analysis, 0.50 for five regression analyses, and 0.49 for nine regression analyses. The ratios corresponding to the number of specified regression analyses were then used to convert Parklawn CPU time to the NIH scale for the purposes of comparison.

A summary of the computation time utilized by the study procedures, for the three distinct production runs and data sets of varying size, is presented in Table 4. As can be observed, the SURREGR procedure was consistently superior in terms of computational efficiency. The SUPERCARP procedure was consistently more efficient than REPERR for all production runs on the data set with 1 120 observations. This is primarily a function of the type of variance estimation algorithm that is employed by the respective

Table 4. Comparison of computation time

Number of		Statistical software procedure			
regression analyses	*n*	SURREGR	SUPERCARP	REPERR[1]	NASSREG
		Central processing unit (CPU) time in seconds			
1	1 120	0.86	2.53	35.58	2.14
1	8 350	3.90	17.60	[2]	14.02
1	38 815	16.59	80.89	47.70	65.69
5	1 120	1.67	7.76	73.63	10.29
5	8 350	5.90	53.45	[2]	69.69
5	38 815	23.45	247.41	86.96	326.96
9	1 120	2.40	13.14	91.51	18.21
9	8 350	7.76	92.10	[2]	123.49
9	38 815	30.50	424.60	114.25	571.19

[1] Converted to NIH CPU time.

[2] Excluded due to the modest variation in CPU time for the smallest and largest data sets.

Source: National Center for Health Services Research and Health Care Technology Assessment.

programs. As noted, REPERR considers a replication approach to variance estimation for complex survey data. This approach identifies a fixed set of representative half-samples from the specified sample (72 for the NMCES design), and requires the computation of regression coefficients for each half-sample.

Although the NASSREG procedure also employs a replication approach to variance estimation, it was also consistently more efficient than REPERR for all production runs on the data set with 1 120 observations. This difference in computational efficiency is due in part to the half-sample weight derivation that occurs in REPERR. As noted, the NASSREG procedure assumes user specification of the set of half-sample replicate weights required for variance computation. The independent derivation of these weights requires use of an orthogonal matrix to define the structure of the half-samples on the data set. The matrix is usually constructed using the technique developed by Plackett and Burman (1943). The number of half-samples is a function of the number of strata employed in the study sample design, equal to the smallest multiple of four which is equal to or greater than the number of strata. Consequently, the NMCES sample design necessitated the computation of 72 sets of half-sample replicate weights for purposes of variance estimation.

When attention is directed to the CPU time used for the production runs on the data set that characterized the entire NMCES sample of 38 815 observations, the REPERR procedure is consistently more efficient than SUPERCARP. Unlike SURREGR and SUPERCARP, where processing time is more directly affected by data base sample size, the processing time for the REPERR procedure is primarily driven by the required recomputation of regression coefficients for a fixed set of specified half-samples. More specifically, there is a point of intersection in the computer processing time functions that characterize SUPERCARP and REPERR, where the inefficiency of REPERR due to the fixed number of replicated computations is outweighed by the greater sensitivity of SUPERCARP to data base sample size. NASSREG is generally the least efficient procedure, adversely affected by the required recomputation of regression coefficients for the specified set of half-samples, and sensitive to data base size.

To determine whether economies of scale were achieved when the data base sample size was increased, a comparison of the CPU time required per observation was considered. Controlling for the specified number of regression analyses, a pattern of greater program efficiency on a relative scale was detected for the larger data sets (Table 5). The decrease in CPU time per observation was generally much greater for the comparisons of data sets of respective sizes 1 120 and 8 350. A much more modest decrease was noted when the comparisons focused on data sets of sizes 8 350 and 38 815, indicating potential convergence to a fixed level for these large data sets. The REPERR procedure was characterized by the greatest relative decrease in CPU time per observation for increasing levels of sample size. The SURREGR program achieved moderate gains in relative efficiency for the larger data sets, whereas only minimal gains were achieved by the SUPERCARP and NASSREG programs.

Table 5. Economies of scale – CPU time per 1 000 observations

Number of regression analyses	*n*	Statistical software procedure			
		SURREGR	SUPERCARP	REPERR[1]	NASSREG
		CPU time per 1 000 observations			
1	1 120	0.77	2.26	31.76	1.91
1	8 350	0.47	2.11	(2)	1.68
1	38 815	0.43	2.08	1.23	1.69
5	1 120	1.49	6.93	65.75	9.19
5	8 350	0.71	6.40	(2)	8.35
5	38 815	0.61	6.38	2.25	8.42
9	1 120	2.14	11.74	81.72	16.26
9	8 350	0.93	11.03	(2)	14.79
9	38 815	0.78	10.94	2.97	14.72

[1] Converted to NIH CPU time.
[2] Excluded due to the modest variation in CPU time for the smallest and largest data sets.

Source: National Center for Health Services Research and Health Care Technology Assessment.

Additional economies of scale were noted when the number of specified regression analyses were varied. Controlling for sample size, the CPU time required per regression analysis consistently decreased as the number of specified regression analyses were incremented (Table 6). This relationship was most noticeable for the REPERR and SURREGR procedures, which experienced the greatest relative efficiencies.

Table 6. Economies of scale – CPU time per regression analysis

Number of regression analyses	*n*	Statistical software procedure SURREGR	SUPERCARP	REPERR[1]	NASSREG
		CPU time per regression analysis			
1	1 120	0.86	2.53	35.58	2.14
1	8 350	3.90	17.60	(2)	14.02
1	38 815	16.59	80.89	47.70	65.69
5	1 120	0.33	1.55	14.73	2.06
5	8 350	1.18	10.69	(2)	13.94
5	38 815	4.69	49.48	17.39	65.39
9	1 120	0.27	1.46	10.17	2.02
9	8 350	0.86	10.23	(2)	13.72
9	38 815	3.39	47.18	12.69	63.47

[1] Converted to NIH CPU time.
[2] Excluded due to the modest variation in CPU time for the smallest and largest data sets.

Source: National Center for Health Services Research and Health Care Technology Assessment.

4.3. Program output capabilities

In addition to producing regression coefficient estimates and their standard errors, and performing tests of significance for the individual model coefficients and overall model specification, these programs possess unique capabilities that distinguish them. A summary of program output capabilities is provided in Table 7. The SURREGR procedure includes as program output an estimated variance-covariance matrix for model coefficients, the weighted sample means for all variables specified in the regression model, and the multiple correlation coefficient for each regression specification. SURREGR has the capacity to yield design effects for each regression coefficient, to simultaneously include continuous and categorical effects in a model specification, and to specify a model with no intercept term. As a SAS procedure, it allows for the production of an output data set of estimated regression coefficients, the rows of the related variance-covariance matrix, and the design effects for model coefficients. It also allows for the output of model residuals to a SAS data set.

Table 7. Comparison of program output capabilities

Feature	SURREGR	SUPERCARP	REPERR	NASSREG
Host system	SAS	Stand alone	OSIRIS	SAS
Means of regression variables	*	*	*	
Variances of regression variables		*	*	
Variance-covariance matrix for the regression variables		*		
Estimated regression coefficients	*	*	*	*
Standard errors (or variances) of estimated regression coefficients	*	*	*	*
T or *F*-test for individual regression coefficients	*	*	*	*
R-squared term	*	*	*	*
Variance-covariance matrix for coefficients	*	*		*
Test of the overall fit of the model	*	*	*	*
Contrasts of regression coefficients		*		*
Partial correlation and standardized regression coefficients			*	
Design effects	*		*	
Automatic collapsing of strata		*		
Errors in variables analysis		*		
No intercept model	*	*		*
Output files	*		*	*

Note: * means that the capability is available.

SUPERCARP program output includes the sample size, weighted sample means for all variables included in the regression analysis, and the related variance-covariance matrix. The SUPERCARP procedure accomodates a multistage sample design, allows for finite population correction factors in variance calculations, and provides for automatic collapsing of strata that contain only one primary sampling unit. Furthermore, the program has the flexibility to compute tests of hypotheses for any subsets of the regression parameters, and includes an errors in variables regression procedure option. In addition, the SUPERCARP package allows for standard error computations of statistics expressed in terms of means, totals, and ratios using the Taylor series linearization method, has the capacity for the testing of hypotheses for domain contrasts, and provides for tests of independence in a two-way table.

The REPERR procedure includes as program output the sample size, the weighted mean, the weighted total, and the standard deviation for each variable incorporated in the regression analysis. Program output also includes the sum of squares by variable and sums of cross products between specified variables in the regression analysis. Replications for variance estimation are created by one of the following three methods: balanced half-sample, jackknife, or user specified replications. Additional optional regression summary statistics include: a multiple correlation coefficient (adjusted and unadjusted), an *R*-squared term for model fit (adjusted and unadjusted), partial correlation coefficients, standardized regression coefficients, and design effects for estimated model parameters. Another REPERR feature permits program output to be stored in machine-readable form in a separate file for later processing.

NASSREG program output includes the number of replicates processed, the sample size and related weighted population estimate, and a summary *R*-squared term for model fit. For a given regression model, NASSREG will automatically provide a test for the overall fit of the model. Additional tests for the significance of a subset or linear combination of model coefficients can be specified by the user. As a SAS procedure, NASSREG permits the creation of an output data set consisting of the estimated model parameters for the full sample and for each replicate, in addition to the related variance-covariance matrix.

4.4. Computational accuracy

The four regression programs were further scrutinized in terms of estimated regression coefficients and related standard errors. All four procedures used the method of weighted least squares to estimate model coefficients. Consequently, the observed equivalence across programs for parameter estimates of regression coefficients was expected. An example of this convergence in estimated regression coefficients is presented in Table 8. The model specification for this comparison considered expenditures for outpatient physician contacts as the dependent variable, and the data base consisted of the overall NMCES sample of 38 815 individuals.

As indicated, the SURREGR and SUPERCARP procedures compute standard errors of regression coefficients using the Taylor series linearization method, while REPERR and NASSREG employ replication techniques. The similarity in behavior of the replication and linearization methods of variance estimation derived from large samples has been demonstrated (Kish and Frankel (1974)). A comparison across regression procedures revealed general convergence in standard error estimates. The complementary standard errors for the estimated regression coefficients are also presented in Table 8.

The greatest concordance in standard error estimates was achieved by the SURREGR and SUPERCARP procedures. Although the procedures employing replication techniques for variance estimation exhibited less convergence, this was due in part to the alternative variance estimators employed by REPERR and NASSREG. Variance estimates of regression coefficients are derived in NASSREG by an average of the squared deviations of the half-sample regression coefficient estimates from the full sample estimate, over the specified number of replications. Alternatively, REPERR considers the average of

Table 8. Comparison of estimated regression coefficients and related standard errors

Independent variables	Estimated standard errors of estimated regression coefficients							
	SURREGR	SUPERCARP	REPERR	NASSREG	SURREGR	SUPERCARP	REPERR	NASSREG
Intercept	86.5654	86.5654	86.5655	86.5654	9.1682	9 1697	*	9.3552
Sex	–18.5145	–18.5145	–18.5145	–18.5145	2.6101	2.6105	2.6128	2.6540
Age	1.2738	1.2728	1.2738	1.2738	.0651	.0651	.0653	.0647
Health status	–57.6026	–57.6026	–57.6026	–57.6026	4.9444	4.9452	4.9582	5.0521
Family income	.0002	.0002	.0002	.0002	.0001	.0001	.0001	.0001
SMSA status	23.8715	23.8715	23.8715	23.8715	3.7395	3.7401	3.7991	3.6934
REGION 1	–26.6823	–26.6823	–26.6823	–26.6823	5.0702	5.0711	5.1146	5.1904
REGION 2	–36.0732	–36.0732	–36.0732	–26.0732	4.8941	4.8949	4.9415	4.9813
REGION 3	–30.9027	–30.9027	–30.9027	–30.9027	5.5317	5.5327	5.6134	5.6347
Race	18.3870	18.3870	18.3869	18.3870	3.9326	3.9333	3.9853	3.9568
Education of household head	13.7831	13.7831	13.7831	13.7831	3.1505	3.1510	3.1768	3.2134
Health insurance coverage	–21.3609	–21.3609	–21.3610	–21.3609	7.6999	7.7012	7.8054	7.9415
Private insurance coverage	14.3403	14.3403	14.3403	14.3403	7.3139	7.3152	7.4041	7.3975
Medicaid coverage	42.5595	42.5595	42.5594	42.5595	7.5640	7.5653	7.6548	7.7116

* Not presented in program output.

Note: The dependent variable for this model specification was expenditures for outpatient physician contacts. The data set consisted of the entire NMCES sample of 38 815 individuals. More detail on the variable definitions and coding is provided in Table 1.

squared deviations of each half-sample regression coefficient estimate from its corresponding complementary half-sample estimate (Cohen and Kalsbeek (1981)).

5. Summary

Using data from the National Medical Care Expenditure Survey, four widely used regression analysis programs appropriate for complex survey data were compared. The programs under investigation included: SURREGR, SUPERCARP, REPERR, and NASSREG. The comparison concentrated on user facility, computational efficiency, and analytical flexibility. The study was also designed to measure the effect of alternative specifications for data base size and number of regression analyses on program performance.

This software comparison was directed to variance estimation tasks associated with the economic, sociological, and behavioral multivariate analyses conducted with NMCES data. A representative set of health care utilization, expenditures and morbidity measures were specified as criterion variables, with demographic, economic, and health insurance coverage measures included as predictors in the regression models under consideration. All four programs had the capacity to produce regression coefficient estimates, generate the related variances or standard errors of the coefficient estimates, and perform tests of significance for the individual model coefficients and the overall model specification. The programs, however, possess other unique capabilities that distinguish them.

As a consequence of the frequency of application of multivariate regression analysis for NMCES analytical reports, the identification and subsequent use of the most efficient software procedure should yield substantial savings in survey costs. The SURREGR procedure is the recommended program of choice for these analyses. The SURREGR procedure was straightforward in its application, requiring the minimum number of programming statements for the full set of analyses under consideration. The SURREGR procedure was also the most efficient program, in terms of CPU time used. This finding was consistent over all specifications of data base size and number of specified regression analyses.

6. References

Chu, A., Mohadjer L., Morganstein, D., and Rhoades, M. (1985): NASSREG (National Accident Sampling System Regression). Internal Westat Report, Rockville, MD.

Cohen, S.B., Burt V.L., and Jones, G.K. (1986): Efficiencies in Variance Estimation for Complex Survey Data. The American Statistician, 40, pp. 157–164.

Cohen, S.B. (1983): Present Limitations in the Availability of Statistical Software for the Analysis of Complex Survey Data. Review of Public Data Use, 11, pp. 338–344.

Cohen, S.B. and Kalsbeek, W.D. (1981): National Medical Care Expenditure Survey: Estimation and Sampling Variances in the Household Survey. National Center for Health Services Research, Instruments and Procedures Series, No. 2, Department of Health and PHS Pub. No. 81–3281. Washington, D.C.

Hidiroglou, M.A., Fuller, W.A., and Hickman, R. (1980): SUPERCARP-Sixth Edition. Survey Section, Iowa State University, Ames, IA.

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.

Holt, M.M. (1982): SURREGR: Standard Errors of Regression Coefficients From Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC.

Jones, G.K. (1983): HESBRR (HES Variance and Crosstabulation Program), Version 3. Internal NCHS Report, Hyattsville, MD.

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

Newhouse, J.P. and Phelps, C.E. (1976): New Estimates of Price and Income Elasticities of Medical Care Services. In The Role of Health Insurance in the Health Services Sector, edited by R.N. Rosett, National Bureau of Economic Research, New York.

Plackett, R.L. and Burman, J.P. (1943): The Design of Optimum Multifactorial Experiments. Biometrika, 33, pp. 305–325.

Shah, B.V. (1981a): SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC.

Shah, B.V. (1981b): RATIOEST: Standard Errors Program for Computing Ratio Estimates for Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC.

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

Van Eck, N. (1979): Osiris IV User's Manual: Fifth Edition. Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, MI.

Received November 1986
Revised February 1988